



SACAIR2025

Southern African Conference for
Artificial Intelligence Research

AI at the Edge
Pushing Boundaries | Powering Possibilities

Century City | Western Cape

1 - 5 December 2025

Proceedings of the Southern African Conference for Artificial Intelligence

Volume II

ISBN: 978-1-0370-5280-4 (e-book)

Editors

**Aurona Gerber
Anban W. Pillay**

**Proceedings of the 6th Southern African Conference for
Artificial Intelligence Research**

SACAIR 2025

ISBN: 978-1-0370-5280-4 (e-book)

Editors

Aurona Gerber

Stellenbosch University
Cape Town, South Africa

Anban Pillay

University of KwaZulu-Natal
Durban, South Africa

© The Editors and The Authors

The SACAIR 2025 Proceedings Editors attest as follows:

All conference paper submissions that appear in these proceedings have been double-blind peer reviewed by an international program committee prior to acceptance and publication.

This work is subject to copyright. All rights are solely and exclusively licensed by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed. The advice and information in this book are believed to be true and accurate at the date of publication. While due care has been exercised, technical and electronic anomalies are possible and unavoidable during the compilation process. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made.

Published Online by

*The SACAIR Steering Committee
Private Bag X20, Hatfield, 0028*

December 2025



SACAIR 2025

Online Proceedings of the 6th South African Conference
for Artificial Intelligence Research, SACAIR 2025

ISBN: 978-1-0370-5280-4 (e-book)

Volume II, of the proceedings of SACAIR 2025, the *6th Southern African Conference for Artificial Intelligence Research*¹.

AI at the Edge: Pushing Boundaries, Powering Possibilities

SACAIR conferences are organised by the the SACAIR steering committee. This, the 6th conference, was hosted by the Department of Computer Science at the University of the Western Cape. The conference was held in Century City, Cape Town, South Africa from 1 to 5 December 2025.

¹ <https://sacair.org.za/>

Preface

Foreword from the Conference Chairs

Dear authors and readers,

It is with great pleasure that we write this foreword to the Proceedings of the sixth Southern African Conference for Artificial Intelligence Research (SACAIR 2025) held in-person and hosted by the University of the Western Cape, Cape Town, South Africa from 1 to 5 December 2025. The programme includes an unconference for students on 1 December 2025 (a student-driven student-empowerment event that includes interaction with industry experts, a student symposium and panel discussions, followed by a day of tutorials on 2 December, and the main conference from 3 to 5 December 2025.

The hosts of SACAIR 2025, the Department of Computer Science at the University of the Western Cape (UWC), is a vibrant centre of teaching, research, and innovation, dedicated to advancing knowledge in Computer Science and its applications to real-world challenges. With a strong focus on academic excellence, collaboration, and societal relevance, the department nurtures future-ready graduates and fosters cutting-edge research across diverse areas such as applied data science and artificial intelligence, cybersecurity and digital inclusion, augmented and virtual reality, and state-of-the-art assistive technologies. Its graduates have gone on to hold high-profile positions in leading technology companies — including Google, Amazon, Openserve, all major South African banks, and a range of other local and international tech organisations — reflecting the department's significant and growing impact and reputation.

The department also proudly hosts the Telkom/Aria Technologies Centre of Excellence in Smart Technologies for the Fourth Industrial Revolution (4IR), which serves as a hub for pioneering research and industry collaboration in emerging digital technologies. Through this initiative, UWC Computer Science contributes to developing solutions that drive technological progress and inclusive growth in South Africa and beyond.

As part of UWC — a university internationally recognised for its commitment to access, equity, and excellence — the department embodies the institution's broader mission to use knowledge as a tool for transformation. Hosting SACAIR 2025 at UWC reflects this shared vision of innovation, collaboration, and impact.

SACAIR 2025 is the sixth international conference focused on Artificial Intelligence hosted by the SACAIR Steering Committee, an affiliate of the Centre for AI Research (CAIR), South Africa. The Centre for AI Research (CAIR)² is a South African distributed research network established in 2011 with the aim of building world-class Artificial Intelligence research capacity in Southern Africa. CAIR conducts foundational, directed and applied research into various aspects

² <https://www.cair.org.za/>

of AI through its various research groups based at higher education institutions in South Africa.

Although still a young conference, SACAIR is quickly establishing itself as a premier artificial intelligence conference in the Southern African region. The sixth conference builds on the success of previous conferences. The inaugural CAIR conference, the Forum for AI Research (FAIR 2019), was held in Cape Town, South Africa, in December 2019. SACAIR 2020 was held in February 2021 after being postponed due to the Covid pandemic and SACAIR 2021 was an online event hosted by the University of KwaZulu-Natal in December 2021. The 2022 conference edition was held in Stellenbosch, Western Cape, SACAIR 2023 was held at the 26 Degrees South venue, Muldersdrift, Gauteng, South Africa, from 4-8 December 2023, and SACAIR 2024 was held at the University of the Free State in Bloemfontein from 4 to 6 December 2024.

We are pleased that SACAIR 2025 continued to enjoy the support of the South African Artificial Intelligence research community. The conference, held under the theme of *AI at the Edge: Pushing Boundaries, Powering Possibilities*, brought together a diverse group of researchers and practitioners. As AI systems evolve, their ability to operate in fast-changing, resource-constrained, and high-impact environments is redefining what's possible. SACAIR 2025 explored how AI is being deployed at the edge—whether in real-time decision-making, autonomous systems, or applications in healthcare, finance, and infrastructure. This theme highlights cutting-edge research driving efficiency, resilience, and breakthroughs in AI's most challenging frontiers.

The conference attracted support from many researchers, as either authors of submitted papers or as reviewers or as members of the international program committee. SACAIR 2025 brought together a diverse group of researchers and practitioners in the fields of Computer Science, Information Systems, Law, and Philosophy of AI. The conference was organised as a multi-track conference that covers broad areas of Artificial Intelligence namely:

- Algorithmic and Data-Driven AI (Computer Science).
- Symbolic AI (Knowledge Representation and Reasoning).
- Socio-technical and human-centred AI (Information Systems).
- Responsible and Ethical AI (Philosophy and Law / Humanities).
- Inter- and trans-disciplinary AI research.

The accepted papers include diverse contributions from all areas of Artificial Intelligence including algorithmic, symbolic and data-driven approaches, socio-technical and human-centred AI, and responsible and ethical AI. The conference series, from inception, has brought together nationally and internationally established and emerging researchers from various disciplines, including Computer Science, Mathematics, Statistics, Information Systems, Philosophy and Law. The conference series maintains a strong focus on cultivating and establishing a network of talented students working in AI from across Africa.

A conference of this nature is not possible without the hard work and contributions of many stakeholders. We extend our sincere gratitude to our sponsors: the Artificial Intelligence Journal (AIJ), the National Institute of Computational

Sciences (NITheCS), the Centre for Artificial Intelligence Research (CAIR), the University of the Western Cape and the National Research Foundation (NRF) of South Africa. These sponsors have made it possible to offer generous scholarships to students and emerging academics to participate in the conference. We sincerely thank the technical chairs for their work in overseeing the onerous technical aspects of the conference and the publication of the two volumes of the proceedings, the track chairs for ensuring the integrity of the review process and, importantly, the international panel of reviewers. We gratefully acknowledge the contributions of our keynotes, industry panel participants, authors, and participants. Finally, we extend our gratitude to the the various Local Organising Committee chairs (Unconference, Tutorials and Hackathon) for overseeing various respective aspects of the conference, the local organising committee, the student organisers, and the conference organiser for their substantive contributions to the success of SACAIR 2025.

We trust that you find value in the various contributions compiled in this proceedings.

Mehrdad Ghaziasgar and Louise Leenen
Organising Chairs: SACAIR 2025

December 2025

Message from the Technical Chairs

Dear readers,

This proceedings volume (ISBN: 978-1-0370-5280-4 (e-book)), contains the revised accepted papers of the South African Conference for Artificial Intelligence, 2025 (SACAIR 2025). The proceedings are published in two volumes. The first volume is published in volume 2784 of the Springer Communications in Computer and Information Science (CCIS) series. This volume, the online proceedings, contains revised papers that were not accepted for the Springer volume but deemed of sufficient quality for presentation at the conference. This volume is available online at the SACAIR 2025 website.

The inter- and trans-disciplinary nature of the SACAIR series of conferences in Artificial Intelligence is unique in that it provides a venue for researchers from a diverse set of disciplines, including Computer Science, Engineering, Information Systems, Law, Philosophy, and the Humanities. The organisation of such a conference has to carefully consider the differing research methods, interests, publication standards, and cultures of these disciplines. The conference was thus organised around the following tracks:

- Algorithmic and Data-Driven AI (Computer Science (CS)):
 - Deep Learning and Machine Learning.
 - Computer Vision and Image Processing.
 - Pattern Recognition
 - Evolutionary Computing
 - Biometrics and Cybersecurity in AI
 - Explainable AI
- Symbolic AI (Knowledge Representation and Reasoning (KRR)):
 - Knowledge Representation and Reasoning.
 - Ontologies.
 - Computational Logic.
 - Multi-agent Systems.
 - Agent-based Modelling.
 - Quantum Artificial Intelligence.
- Socio-technical and human-centred AI (Information Systems (IS)):
 - AI Information Systems.
 - AI Socio-Technical Systems.
 - Human-Centred Artificial Intelligence.
 - AI for and in Business, including AI adoption.
 - AI Supporting Sustainable Development and the Circular Economy.
 - AI for Development and Social Good.
- Responsible and Ethical AI (Philosophy and Law (P&L)):
 - Data Ethics.
 - Machine Ethics.
 - Ethics of Socio-robotics.
 - Neuro Ethics.
 - AI and the Law.

- Responsible AI Governance.

The program committee comprised 150 members, representing some 95 research institutions from 27 countries across the globe. Each paper was sent for review to at least three members of the program committee. The technical committee ensured a rigorous, double-blind peer-review process. Most papers received at least three reviews, with a mean of 2.26 per paper. Several meta-reviews were carried out by the respective track chairs. Great care was taken to ensure the conference's integrity, including careful attention to avoid conflicts of interest. The following criteria were used to rate submissions and to guide decisions: relevance to SACAIR, significance, technical quality, scholarship, and presentation (including quality and clarity of writing).

We received more than 150 initial submissions, and after a first round of evaluation, 104 submissions were sent to our SACAIR program committee for review. The papers consisted of 53 in the CS track, 25 in the IS track, 1 in the KRR track and 24 in the P&L track. Thirty-seven full research papers were selected for publication in the Springer CCIS volume (which translates to an acceptance rate of 35.9%). In total, three papers from the Responsible and Ethical AI track, nine papers from the Socio-technical and Human-Centred AI track, and twenty-five papers from the Algorithmic and Data-Driven AI track were accepted for publication in the CCIS volume.

Twenty-two papers were selected for publication in this volume, the online proceedings; seven in the CS track, eight in the IS track, one in the KRR track, and 6 in the P&L track. This translates to an acceptance rate of 21.4% of reviewed submissions and an overall acceptance rate of 57.3% for the two volumes.

Thank you to all the authors who submitted work of an exceptional standard to the conference and congratulations to the authors whose work was accepted for publication. We place on record our gratitude to the Program Committee members, whose thoughtful and constructive comments were well received by the authors and whose diligence ensures the continued high quality of the conference.

Aurona Gerber and Anban Pillay
Technical Chairs: SACAIR 2025

December 2025



November 26, 2025

The Technical Committee
6th Annual conference (SAICAIR 2025)

Attestation: 6th Annual Southern African Conference for Artificial Intelligence Research (SACAIR 2025)

The technical committee of SACAIR 2025 attest as follows: **All conference paper submissions that appear in the online proceedings have been double-blind peer reviewed by an international program committee (PC) prior to acceptance and publication. More than 75% of the articles were authored by researchers from different institutions. The authors represented 96 institutions, with 26,2% of these from outside South Africa**

The 6th Annual Southern African Conference for Artificial Intelligence Research (SACAIR 2025) was hosted by the Department of Computer Science, University of the Western Cape and held in Cape Town, South Africa, from 1 to 5 December, 2025. Submissions were sought in one of five tracks:

- Algorithmic and Data-Driven AI (Computer Science).
- Symbolic AI (Knowledge Representation and Reasoning).
- Socio-technical and human-centred AI (Information Systems).
- Responsible and Ethical AI (Philosophy and Law / Humanities).
- Inter- and trans-disciplinary AI research.

One hundred and forty-nine abstracts were submitted, and after full paper submissions and a first round of evaluation, 104 submissions were sent to the PC for review. Great care was taken to avoid conflicts of interest when assigning reviewers. The review process was double-blind and rigorous, with every paper sent to at least three reviewers and each receiving at least two substantive reviews and most receiving three reviews, with a mean per paper of 2.6%. The program committee comprised 150 members, representing some 95 research institutions from 27 countries across the globe. The following criteria were used to rate submissions and to guide decisions: relevance to SACAIR, significance, technical quality, scholarship, and presentation, which included quality and clarity of writing. The technical committee, comprising the technical chairs and track chairs, managed the review process and supported the authors in revising their papers to the quality required for publication in these proceedings.

Revised, accepted papers were published in two proceedings volumes: Volume 2784 in the conference series [Southern African Conference for Artificial Intelligence Research](#) of the book series [CCIS: Communications in Computer and Information Science](#) and an online proceedings (ISBN: 978-1-0370-5280-4 (e-book)) that included papers that were not accepted into Volume I but were of sufficient quality to be presented at the conference and published in this volume.

The acceptance rate for full research papers for the CCIS volume was 35.9%. The acceptance rates for reviewed papers in the different tracks were: 36.0% in the Information Systems track, 47.2% in the Computer Science track, and 12.5% for the Philosophy and Law track. In addition, and based on reviews, 22 papers were accepted into Volume II of the proceedings. The acceptance rate for Information Systems papers for Volume II of the proceedings was 32.0%, 13.2% for the Computer Science track and 25.0% for the Philosophy and Law track. The overall acceptance rate for the conference (for both volumes) was thus 57.3%. Table 1 gives the final submission and acceptance data for the conference.

Track	Subs	Review	Springer	Online	Reject	Springer%	Online%	Total%
SACAIR - CS	83	53	25	7	21	47.2%	13.2%	60.4%
SACAIR - ID	1	0	0	0	0	0.0%	0.0%	0.0%
SACAIR - IS	30	25	9	8	8	36.0%	32.0%	68.0%
SACAIR - KRR	2	1	0	1	0	0.0%	100.0%	100.0%
SACAIR - P&L	33	24	3	6	15	12.5%	25.0%	37.5%
TOTAL	149	103	37	22	44	35.9%	21.4%	57.3%

Table 1: Submission statistics by track with acceptance rates based on reviewed submissions.

Authors of accepted papers were affiliated with various national and international universities. Table 2 below indicates the percentage of authors from the respective institutions for the online proceedings volume. The authors represented 64 national and international institutions. We thus confirm that more than 75% of the papers were authored by researchers from different universities.

Institution	Count	Percentage
University of Pretoria	14	9.3%
University of Cape Town	13	8.7%
University of Johannesburg	10	6.7%
University of South Africa	9	6.0%
North-West University	9	6.0%
University of the Witwatersrand	8	5.3%
Stellenbosch University	6	4.0%
FHNW University of Applied Sciences and Arts No...	6	4.0%
University of KwaZulu-Natal	5	3.3%
University of Edinburgh	4	2.7%
Other	66	44.0%

Table 2: Top 10 institutions in PC composition (all tracks). Total unique institutions is 64.

The Technical Committee
SAICSIT • 2025
November 26, 2025

Our Sponsors

The generous support of our sponsors is gratefully acknowledged.



UNIVERSITY *of the*
WESTERN CAPE



National
Research
Foundation

NITheCS

National Institute for
Theoretical and Computational Sciences

Local Organizing Committee

Local Organising Committee

General Co-Chairs

Prof Mehrdad Ghaziasgar
Prof Louise Leenen

University of the Western Cape
University of the Western Cape

Unconference Co-Chairs

Dr Oluwakorede M. Oluyide
Mr Clayton K. Baker
Mr Mluleki Bryan Sinquadu
Mr Sibonelo Dlamini
Ms Yuvika Singh
Dr Tevin Moodley

University of the Western Cape
University of the Western Cape
University of KwaZulu-Natal
University of KwaZulu-Natal
University of KwaZulu-Natal
University of Johannesburg

Tutorials Chair

Dr Ayinde Mohammed Usman

University of the Western Cape

Program Committee

Algorithmic and Data Driven AI

Track Chairs

Davel, Marelle	North-West University, South Africa
Gerber, Aurora	CAIR, Stellenbosch University, South Africa
Kotzé, Eduan	University of the Free State, South Africa
Leenen, Louise	University of the Western Cape, South Africa
Van Zyl, Terence	University of Johannesburg, South Africa

PC member

Amuda, Kamorudeen	University of Ibadan, Nigeria
Anderson, George	University of Botswana, Botswana
Asaju, Christine	Thswane University of Technology, South Africa
Atemkeng, Marcellin	Rhodes University, South Africa
Barnard, Etienne	North-West University, South Africa
Brown, Dane	Rhodes University, South Africa
Buyse, Jan	University of Cape Town, South Africa
Coulter, Duncan	University of Johannesburg, South Africa
Daramola, Olawande	University of Pretoria, South Africa
De Freitas, Allan	University of Pretoria, South Africa
Derks, Iena	University of Pretoria, South Africa
Dlamini, Sibonelo	University of KwaZulu-Natal, South Africa
Du Toit, Tiny	North-West University, South Africa
Eiselen, Roald	North-West University, South Africa
Ezeji, Ijeoma	University of Zululand, South Africa
Fabris-Rotelli, Inger	University of Pretoria, South Africa
Fashoto, Stephen	University of Swaziland, Eswatini
Haskins, Bertram	Nelson Mandela University, South Africa
Helberg, Albert	North-West University, South Africa
Hounkanrin, Anicet	University of Cape Town, South Africa
Ingram, Branden	University of Witwatersrand, South Africa
Isingizwe, Frederic	University of the Western Cape, South Africa
James, Steven	University of the Witwatersrand, South Africa
Jarvis, Devon	University of the Witwatersrand, South Africa
Jideani, Paul	University of the Western Cape, South Africa
Kankuzi, Bennett	MUST, Malawi
Kanyama, Maria Nelago	Namibia University of Science and Technology, Namibia
Klein, Richard	University of the Witwatersrand, South Africa

Kogeda, Okuthe Paul	University of KwaZulu-Natal, South Africa
Laidi, Amel	Abou Bekr Belkaïd University, Algeria
Makhoere, Lucky	Sol Plaatje University, South Africa
Malan, Katherine	University of South Africa, South Africa
Marais, Patrick	University of Cape Town, South Africa
Masethe, Dan	Tshwane University of Technology, South Africa
Mathonsi, Topside	Tshwane University of Technology, South Africa
Mauwa, Hope	University of Mpumalanga, South Africa
Melatagia Yonta, Paulin	University of Yaounde I, Cameroon
Modipa, Thipe	University of Limpopo, South Africa
Moges, Henok	University of Cape Town, South Africa
Moodley, Deshendran	University of Cape Town, South Africa
Motsoehli, Moseli	University of Hawaii at Manoa, United States
Mouton, Coenraad	Multilingual Speech Technologies, North-West University, South Africa
Mungloo-Dilmohamud, Zahra	University of Mauritius, Mauritius
Mutembei, Leonard	University of Johannesburg, South Africa
Nandutu, Irene	Uganda Technology and Management University, Uganda
Nangue Tasse, Geraud	University of the Witwatersrand, South Africa
Nemakhavhani, Ndamulelo	University of Pretoria, South Africa
Ngxande, Mkhusele	Stellenbosch University, South Africa
Nicolls, Fred	University of Cape Town, South Africa
Nogwina, Mnoneleli	Sol Plaatje University, South Africa
Nzomo, Mbithe	University of Cape Town, South Africa
Ojino, Ronald	The Open University of Kenya, Kenya
Okuboyejo, Damilola	Durban University Of Technology, South Africa
Okuboyejo, Olaperi	University of the Witwatersrand, South Africa
Olukanmi, Seun	University of Witwatersrand, South Africa
Olusanya, Micheal	Sol Plaatje University Kimberley, South Africa
Omollo, Richard	Jaramogi Oginga Odinga University of Science and Technology, Kenya
Omomule, Taiwo	South African National Bioinformatics Institute, South Africa
Onuegbu, Joseph	Ariel University, Israel
Oriola, Oluwafemi	University of the Free State, South Africa
Rabe, Randle	MUST Deep Learning, NWU, South Africa
Rananga, Seani	University of Pretoria, South Africa
Rens, Gavin	Katholieke Universiteit Leuven, Belgium
Sefara, Tshephisho Joseph	CSIR, South Africa
Semwayo, Thembinkosi	University of the Witwatersrand, Johannesburg, South Africa
Senekane, Makhamisa	University of Johannesburg, South Africa
Shock, Jonathan	University of Cape Town, South Africa
Tachiona, Terence	Midlands State University, Zimbabwe

Tapamo, Jules-Raymond	University of KwaZulu-Natal, South Africa
Theunissen, Tian	MuST North-West University, South Africa
Tom, Sheethal	Cape Peninsula University of Technology, South Africa
Tuyishimire, Emmanuel	University of Cape Town, South Africa
Vadapalli, Hima Bindu	University of Johannesburg, South Africa
Varadarajan, Swetha	University of Cape Town, South Africa
Watson, Bruce	National Security Centre of Excellent (Canada), South Africa

Socio-technical and Human-Centred AI

Track Chairs

Eybers, Sunet	University of South Africa (UNISA), South Africa
Gerber, Aurona	CAIR, Stellenbosch University, South Africa
Hinkelmann, Knut	FHNW University of Applied Sciences and Arts Northwestern Switzerland, Switzerland

PC members

Abdullah, Hanifa	UNISA, South Africa
Botha, Benjamin	University of the Free State, South Africa
Buchmann, Robert Andrei	Babeş-Bolyai University of Cluj Napoca, Romania
Callisto De Donato, Massimo	University of Camerino, Italy
Choshi, Sonti	UNISA, South Africa
Da Veiga, Adele	UNISA, South Africa
Ehrenthal, Joachim	FHNW, Switzerland
Fornari, Fabrizio	University of Camerino, Italy
Hanslo, Ridewaan	University of Johannesburg, South Africa
Hoppe, Thomas	Datenlabor Berlin, Germany
Howard, Grant Royd	University of South Africa (UNISA), South Africa
Härer, Felix	FHNW University of Applied Sciences and Arts, Basel, Switzerland
Kaufmann, Michael Alexander	Lucerne University of Applied Sciences and Arts, Switzerland
Labuschagne, Celeste	Belgium Campus iTVersity, South Africa
Lotriet, Hugo	UNISA, South Africa
Mahlaza, Zola	University of Cape Town, South Africa
Martin, Andreas	FHNW University of Applied Sciences Northwestern Switzerland, Switzerland
Matthee, Machdel	University of Pretoria, South Africa
Maus, Heiko	German Research Center for AI (DFKI), Germany
Mennega, Nita	University of Pretoria, South Africa

Mentz, Jan	Belgium Campus ITversity, South Africa
Parry, Douglas	Stellenbosch University, South Africa
Pretorius, Henk	University of Pretoria, South Africa
Price, Catherine S.	University of KwaZulu-Natal, South Africa
Pustulka-Hunt, Ela	FHNW Olten, Switzerland
Scholtz, Dorothy	University of South Africa, South Africa
Sibanyoni, Nhlanhla	University of Pretoria, South Africa
Smuts, Hanlie	University of Pretoria, Pretoria, South Africa, South Africa
Telesko, Rainer	University of Applied Sciences Northwestern Switzerland, Switzerland
Tiddi, Ilaria	Vrije Universiteit Amsterdam, Netherlands
Van Belle, Jean-Paul	University of Cape Town, South Africa
Van Elst, Ludger	DFKI, Germany
Van Staden, Corné	UNISA, South Africa
Watson, Bruce	National Security Centre of Excellent (Canada), South Africa
Weilbach, Lizette	University of Pretoria, South Africa
Witschel, Hans Friedrich	Fachhochschule Nordwestschweiz, Switzerland

Responsible and Ethical AI

Track Chairs

Benton, Paige	University of Johannesburg, South Africa
Ruttkamp-Bloem, Emma	Department of Philosophy, University of Pretoria, South Africa
Tollon, Fabio	University of Edinburgh, South Africa

PC members

Basu, Kaberi	Trinity College, Dublin, Ireland
Bock, Ivan	Stellenbosch University, South Africa
Brand, Joshua	Institut Polytechnique de Paris (Télécom Paris), France
Chimba, Bridget	University of Pretoria, South Africa
Ching-Lam, Janice	Chinese University of Hong Kong), Hong Kong
Claassen, Kristy	University of Twente, Netherlands
Ferdman, Avigail	Department of Humanities and Arts, Technion - Israel Institute of Technology, Israel
Frost-Arnold, Karen	Hobart & William Smith Colleges, United States
Galvagni, Enrico	University of Edinburgh, United Kingdom
Garforth, James	The University of Edinburgh, United Kingdom
Howdle, Giles	Utrecht University, Netherlands

Miragoli, Martin	University of Johannesburg, South Africa
Moletsane, Dimpho	University of Johannesburg, South Africa
Poyatos, Juanma	University of Auckland, New Zealand
Rebera, Andrew	AND Consulting Group, Belgium
Schulz, Maria	Tilburg University, Germany
Smit, Sasha Lee	University of Edinburgh, United Kingdom
Timcke, Scott	Research ICT Africa, South Africa
Townsend, Bev	University of York, United Kingdom
Vorster, Ethan	University of Johannesburg, South Africa
de Villiers-Botha, Tanya	Stellenbosch University, South Africa
van Iersel, Nanou	Erasmus University Rotterdam, Netherlands
Önder, Orhan	Marmara University, Turkey

SACAIR - KRR

Track Chairs

Casini, Giovanni	CNR - ISTI, Italy
Heyninck, Jesse	Open Universiteit, the Netherlands, Germany

PC members

Booth, Richard	Cardiff University, United Kingdom
Haldimann, Jonas Philipp	TU Wien, Austria
Leenen, Louise	University of the Western Cape, South Africa
Meyer, Thomas	University of Cape Town and CAIR, South Africa

Table of Contents

I Volume I: Algorithmic and Data Driven AI	1
Which contextual topic modelling algorithm is the best?	3
<i>Darren Roos and Katherine Malan</i>	
Leveraging Language Models for Document Type Classification in Low-Resource Afrikaans Archives	4
<i>Eduan Kotzé and Burgert Senekal and Walter Daelemans</i>	
Hybrid Automatic Modulation Classification for Increased Robustness under White-Box Adversarial Attacks	5
<i>Arnold van der Merwe and Albert Helberg</i>	
Counterfactual Explanation Model for Personalised Dietary Interventions in Anaemia Patients	6
<i>Zvinodashe Revesai and Okuthe P. Kogeda</i>	
Feature extraction for plant growth estimation	7
<i>Simbarashe Aldrin Ngorima and Albert Helberg and Marelie Davel</i>	
Exploring the impact of different loss functions for anomaly prediction in a mineral processing plant	8
<i>Morne C. Du Plessis and Deshendran Moodley</i>	
Modifying class distributions to improve the classification of minority group examples in a class-imbalanced dataset	9
<i>Banele Mdluli and Terence Van Zyl</i>	
Evaluating OCR performance on food packaging labels in South Africa	10
<i>Mayimunah Nagayi and Clement Nyirenda and Rina Swart and Alice Khan and Tamryn Frank</i>	
Investigating the Impact of Multilingual Pre-trained Speech Models on Gender Bias in ASR for Low Resource African Languages	11
<i>Claytone Sikasote and Hussein Suleman and Jan Buys</i>	

A Two Stage Pipeline for Automated Caries Detection on Single Tooth Images from Panoramic Radiographs	12
<i>Christopher Jürgen Hansen and Coenraad Mouton and Claus-Christian Glüer and Paula Klöhn and Anna-Louisa Kollster and Christof Dörfer and Jonas Conrad and Christian Graetz and Toni Gehrmann and Niklas Christoph Koser and Jan-Bernd Hövener</i>	
Exploring syllable similarity across South African languages through self-supervised speech representation	13
<i>Aby Louw</i>	
Data Pruning: Redundant, Problematic, and Interdependent Samples	14
<i>Leon Freese and Marthinus Wilhelmus Theunissen</i>	
Restoring Neural Network Plasticity for Faster Transfer Learning	15
<i>Xander Coetzer and Anna Bosman and Arné Schreuder</i>	
AfroXLMR-Comet: Multilingual Knowledge Distillation with Attention Matching for Low-Resource Languages	16
<i>Joshua Sakthivel Raju and Sanjay Somasundaram and Jaskaran Singh Walia and Srinivas R and Vukosi Marivate</i>	
Code-Switch Pretraining for Improved Cross-Lingual Alignment in Low-Resource Languages	17
<i>Ruan Visser and Trienko Grobler and Marcel Dunaiski</i>	
Identification of Social Media Users that Perpetuate Xenophobic Attitudes and Hate Speech Narratives in South Africa	18
<i>Carl du Plessis and Michael du Plessis and Ronny Mabokela and Abiodun Modupe and Vukosi Marivate</i>	
HAPI-FedProx: Heterogeneity-Aware Adaptive Proximal Optimization for Federated Learning	19
<i>Kudzaishe Mhou and Senekane Makhamisa</i>	
Fog-Based Deep Learning for Real-Time Cold Chain Temperature Prediction Using IoT Data	20
<i>Jeremiah Taguta and Jean Frederic Isingizwe Nturambirwe and Clement Nthambazale Nyirenda</i>	
Robust Driver Distraction Recognition via Lightweight Body-Part Association and Object Context on NVIDIA Jetson	21
<i>Frank Zandamela and Patrick Malatjie and Teboho Sekopa and Mamodike Sadiki and Moloko Manthata</i>	

Nearest-Class Mean and Logits Agreement for Wildlife Open-Set Recognition	22
<i>Jiahao Huo and Mufhumudzi Muthivhi and Fredrik Gustafsson and Terence Van Zyl</i>	
Semi-Supervised Object Segmentation via Active Learning for Efficient Ecological Monitoring	23
<i>Dane Brown and Karen Bradshaw</i>	
Scaling behavior of Encoder Language Models in Low-Resource Settings	24
<i>Ruan Visser and Trienko Grobler and Marcel Dunaiski</i>	
[Full Paper] Enhancing Target Re-Identification via Model Fusion and Knowledge Distillation of Pre-trained Foundation Models	25
<i>Tendai Shoko and Prof. Terence L. van Zyl</i>	
Heart Disease Prediction: A Comparative Study of Optimizers' Performance in Deep Neural Networks	26
<i>Chisom Chibuike and Adeyinka Ogunsanya</i>	
II Volume I: Socio-technical and Human-Centred AI	27
AI-Enabled Cybersecurity Implementation: A Case Study of Critical Success Factors in a South African State-Owned Entity	29
<i>Awonke Mamane and Rennie Naidoo</i>	
RAG Evaluation: From Model-Centric Benchmarks to System-Level Metrics	30
<i>Alta de Waal and Daniel Van Niekerk and Florian Donhauser and Salmaan Suliman and Dehan Lamprecht</i>	
Exploring Postgraduate Students' Use and Perceptions of Generative AI as a Research Support Tool: A South African Case	31
<i>Mfowabo Maphosa and Lucas Khoza and Cyril Tlomatsana and Winnie Pitjo</i>	
An Awareness Framework for Sustainable Selection of LLMs in Business	32
<i>Natascha Brughitta Anchia and Andreas Martin</i>	
Synthesizing Insights on AI in Africa: A Comprehensive Qualitative Analysis	33
<i>Arti Ramanathan and Dongpeng Huang and James Katz</i>	

Factors influencing the use of Autonomous Vehicles on Traditional Road Infrastructure in The Western Cape	34
<i>Naasir Parker and Walter Uys</i>	
Large Language Model Powered Agentic Framework for Cholera Risk Prediction with Explainable ML and Statistical Insights	35
<i>Paul Jideani and Aurona Gerber</i>	
A Scoping Review on Generative AI prompting to optimize the workflow of healthcare professionals in sub-Saharan Africa	36
<i>Oluwamayowa O. Ogundaini and Lisa-Dionne Morris</i>	
Detection Of Power Line Insulator Defects Using YOLOv10-N	37
<i>Kgampu Shawn Shawn and Terence van Zyl</i>	
III Volume I: Responsible and Ethical AI	39
Artificially Generated Minorities (AGMs): The Veneer of Algorithmic Bias Correction	41
<i>Ibifuro R. Jaja</i>	
Under-resourced and over-researched: The ELSA approach to transdisciplinary AI research in low-trust neighborhoods	42
<i>Nanou van Iersel and Majsa Storbeck and Marlon Kruizinga and Michaël Grauwde</i>	
From Principles to Practice: Extracting Human-Centred AI Components from Governance Frameworks	43
<i>Cindy van den Berg and Hanlie Smuts</i>	
IV Volume II: Algorithmic and Data Driven AI	45
MAGE: Multi-Head Attention Guided Embeddings for Low Resource Sentiment Classification	47
<i>Varun Vashisht and Samar Singh and Mihir Konduskar and Jaskaran Singh Walia and Vukosi Marivate</i>	
An End-to-End Deep Learning Model for Recommender Systems	64
<i>Abebe Tegene and Vukosi Marivate and Mapundi Banda and Abiodun Modupe and Valisoa Rakotonarivo and Mathibele Nchabeleng</i>	
Should We Trust AI to Detect Social Stress? A Machine Learning Approach	82
<i>Valentina Oelofse and Herkulaas Combrink</i>	

Cross-lingual transfer of multilingual models on low resource African Languages	94
<i>Harish Thangaraj and Ananya Chenat and Jaskaran Singh Walia and Vukosi Marivate</i>	
Injecting Explicit Cross-lingual Embeddings into Pre-trained Multilingual Models for Code-Switching Detection	104
<i>Thapelo Sindane and Vukosi Marivate and Avashlin Moodley</i>	
V Volume II: Socio-technical and Human-Centred AI	123
Benchmarking and deploying Small Language Models on the edge for real-world agentic systems in industry	125
<i>Edward Duffy and David Fernandez and Alta de Waal and Mert Pesé</i>	
Evaluating the Backup Buddy Chatbot for Raising Awareness of Mobile Bullying	140
<i>Portia Mathimbi and Wallace Chigona</i>	
Using AI to boost scoping reviews. Exploring AI deployment in obstetrics and gynaecology as an exemplar	162
<i>Isabella Venter and Renette Blignaut and Karen Renaud</i>	
A Cross-Cultural Assessment of Human Ability to Detect LLM-Generated Fake News about South Africa	174
<i>Tim Schlippe and Matthias Wölfel and Koena Mabokela</i>	
Persona-Aware Alignment of LLMs Using Synthetic Dialogue Data	191
<i>Annick Grob and Hans Friedrich Witschel and Andreas Martin</i>	
R.A.I.S.E - A Novel Framework for Evaluating Foundational AI Models in Medical Deployment: Moving Beyond Traditional Metrics to Real-World Deployability	208
<i>Zardus Adendorff and Laing Lourens and Judy Gichoya and Vukosi Marivate and Rhena Delpot</i>	
A perspective on Agentic AI as a component of the analytics workflow	224
<i>Charl Cowley and Warren Brettenny</i>	
The Us es of Generative Artificial Intelligence for Cybersecurity in Organisations	244
<i>Chipo Chidakwa and Zainab Ruhwanya</i>	

VI Volume II: Responsible and Ethical AI	257
Otheroids or Anthropomorphism? An Empathy-Based Approach to Artificial Agents	259
<i>Abootaleb Safdari</i>	
Bestowing Legal Personality on AI: A South African Perspective	268
<i>Kristi Erasmus</i>	
Digital Colonialism as an Economic Strategy: Engineered In- equality	292
<i>Sunita Menon</i>	
"Where you want her, how you want her": Understanding the violence of deepfake pornography	307
<i>Amukelani Mnisi</i>	
An initial foray into AI judicial drafting: a comparative exper- iment in a constitutional privacy case.	326
<i>Kudakwashe Mawowa and Donrich Thaldar</i>	
 VII Volume II: SACAIR - KRR	347
An Override-Aware Classifier for Transparent AI	349
<i>Ruvarashe Shalom Madzime and Thomas Meyer and Louise Leenen</i>	

Part I

**Volume I: Algorithmic and
Data Driven AI**

Which contextual topic modelling algorithm is best?

Darren Craig Roos¹[0000–0002–0405–524X] and Katherine Mary
Malan¹[0000–0002–6070–2632]

Department of Decision Sciences, University of South Africa, Pretoria, South Africa
21001154@mylife.unisa.ac.za

Abstract. The question of which contextual topic modelling algorithm performs best has become increasingly important as the field rapidly develops new approaches. However, existing evaluations typically focus on limited datasets and metrics, often claiming superiority for novel algorithms. This study presents a comprehensive empirical evaluation of eleven contextual topic modelling algorithms across ten diverse datasets, five numbers of topics, and four performance metrics, resulting in 22,000 metric evaluations. Rather than identifying a single superior algorithm, our results reveal clear evidence of performance complementarity: different algorithms excel on different problem instances and under different evaluation criteria. Through aggregate performance analysis, pairwise dominance comparisons, and multi-objective Pareto frontier analysis, we demonstrate that algorithmic dominance varies significantly across problem instances. Most remarkably, in 84% of cases, all algorithms are Pareto optimal when considering all metrics simultaneously, indicating that each offers unique strengths that cannot be dominated by others. These findings challenge the common practice of claiming algorithmic superiority and suggest that algorithm selection should be guided by specific problem characteristics and performance priorities rather than blanket recommendations. Our work contributes to the growing recognition that performance complementarity is fundamental to computational problems, extending this concept to contextual topic modelling and providing a foundation for future algorithm selection research. Code used to conduct this study is provided.¹

Keywords: Natural language processing · Topic modelling · Performance analysis · Performance complementarity.

¹ https://github.com/AlgorithmicAmoeba/tm_framework

Leveraging Language Models for Document Type Classification in Low-Resource Afrikaans Archives

Eduan Kotze¹[0000-0002-5572-4319], Burgert A. Senekal¹[0000-0002-1385-9258],
and Walter Daelemans²[0000-0002-9832-7890]

¹ Department of Computer Science and Informatics, University of the Free State,
Bloemfontein, South Africa
kotzeje@ufs.ac.za, burgertsenekal@yahoo.co.uk
² CLIPS - Computational Linguistics Group, University of Antwerp, Antwerp,
Belgium
walter.daelemans@uantwerpen.be

Abstract. Document type classification is essential for effective information retrieval and management within archival systems, particularly in low-resource languages like Afrikaans. This study examines the feasibility of utilising multilingual transformer-based language models for document classification within a South African archival context. We followed a basic linguistic approach to prepare Afrikaans text documents for classification into six categories: academic papers, media reports, books, interviews, book reviews, and theses or dissertations. We compare fine-tuned transformer models, hybrid models combining traditional classifiers with contextual embeddings, and a baseline SVM (TF-IDF) classifier, using stratified 5-fold cross-validation and a hard voting ensemble for robust evaluation. Our findings reveal that the SERENGETI transformer-based model outperformed other multilingual models, achieving a weighted F1 score of 0.964, while hybrid approaches performed competitively. However, the baseline SVM (TF-IDF) model outperformed all transformer and hybrid models, with a weighted F1 score of 0.978. This research demonstrates the potential and current limitations of neural language models and hybrid strategies for enhancing document classification in Afrikaans archival systems. If implemented, the classifier can improve indexing efforts and reduce pressure on archival personnel who handle over 5,000 new items annually.

Keywords: Document Classification · Archival Systems · BERT · XLM-V · SERENGETI · Hybrid models · Afrikaans · Low-resource languages.

1 Introduction

Archives face numerous challenges in the digital age. The volume of data that needs to be processed and preserved has grown exponentially, making it difficult

Hybrid Automatic Modulation Classification for Increased Robustness under White-Box Adversarial Attacks

A van der Merwe^[0009–0001–5032–7090] and ASJ Helberg^[0000–0001–6833–5163]

Faculty of Engineering, North-West University, Potchefstroom, South Africa

arnold1.vandermerwe@gmail.com

albert.helberg@nwu.ac.za

Abstract. Automatic modulation classification (AMC) is an important function in wireless communication systems that is used to identify the modulation type of a signal without prior knowledge. AMC has historically been done using likelihood or feature-based methods, yet recent research has focused on using deep neural networks (DNNs) as they outperform the classical methods in challenging signal channel conditions. However, deep learning (DL) based classifiers are vulnerable to adversarial attacks that can significantly deteriorate their classification performance. This paper explores the robustness of different AMC classifiers to the white-box fast gradient method (FGM) and projected gradient descent (PGD) attacks under different perturbation-to-noise ratios (PNRs) and signal-to-noise ratios (SNRs) for a noisy signal channel. The investigated AMC classifiers consist of the quasi-hybrid likelihood ratio test (QHLRT), a k-nearest neighbour (KNN) that uses higher-order cumulants, and the parameter estimation and transformation-based CNN-GRU deep neural network (PET-CGDNN). The adversarial attacks are found to have limited transferability to the QHLRT and the KNN classifiers when scaled to be imperceptible against the noise of the signal channel. Based on this finding, we propose a hybrid classifier that uses the neural rejection technique through a support vector machine (SVM) that acts as a switching mechanism to decide whether to use the KNN or PET-CGDNN to classify the modulation type. The hybrid classifier demonstrates improved robustness against attacks, while benefiting from the good performance of the DNN.

Keywords: Modulation classification · Adversarial attacks · Deep neural networks · Likelihood ratio test · K-nearest neighbours

1 Introduction

Artificial intelligence (AI) has increasingly become a foundational technology in a wide range of industries, including telecommunications, through the implementation of deep neural networks (DNNs). These DNNs are used to automate network management, optimise resource allocation, and improve signal detection

Counterfactual Explanation Model for Personalised Dietary Interventions in Anaemia Patients

Zvinodashe Revesai^{1[0009-0008-2284-6097]} and Okuthe P. Kogeda *^[0000-0002-8353-8345]

¹ School of Mathematics, Statistics and Computer Science, College of Agriculture, Engineering and Science, University of KwaZulu-Natal, Westville Campus, Durban 3209, Republic of South Africa.

224195689@stu.ukzn.ac.za, kogedao@ukzn.ac.za*

Abstract. Deep learning has revolutionised healthcare applications, achieving remarkable success in medical diagnosis and treatment prediction. However, the inherent opacity of these models presents significant challenges for clinical deployment, where interpretable explanations are crucial for patient trust and regulatory compliance. This paper presents a novel constraint-aware counterfactual explanation model for generating personalised dietary interventions in anaemia patients. Anaemia affects over 1.9 billion people globally, yet existing explainable AI methods fail to provide clinically feasible and culturally appropriate recommendations. We develop a causal machine learning approach that integrates Pearl's causal hierarchy with domain-specific constraints to produce interpretable "what-if" scenarios. Our model incorporates nutritional, cultural, and economic constraints through augmented Lagrangian optimisation, ensuring recommendations remain clinically feasible whilst maintaining semantic meaningfulness. Experimental results demonstrate superior performance compared to existing explainable AI methods, achieving 84.3% anaemia reversal rates (vs 71.8% best baseline), 89.1% counterfactual validity, and 4.2 interpretability scores. The model generates recommendations requiring an average of 2.3 dietary changes within cognitive load thresholds whilst maintaining $O(n \log n)$ computational complexity suitable for real-time clinical deployment. This work advances explainable AI in healthcare by demonstrating how domain-specific constraints can enhance both interpretability and clinical utility of counterfactual explanations for chronic disease management.

Keywords: Counterfactual Explanations, Explainable AI, Causal Machine Learning, Healthcare Informatics, Interpretability

1 Introduction

Deep learning has revolutionised healthcare applications, achieving remarkable success in medical image analysis, drug discovery, and clinical decision support systems [1]. However, the inherent opacity of these models presents significant challenges for healthcare deployment, where clinical decisions require transparent reasoning and accountability [2]. The black box nature limits adoption in critical scenarios, as medical professionals need to understand algorithmic recommendations to ensure patient safety and regulatory compliance. Consequently, developing interpretable machine learning

Feature extraction for plant growth estimation

SA Ngorima^{1,2,3[0000-0002-0775-3529]}, ASJ Helberg^{1[0000-0001-6833-5163]}, and
MH Davel^{1,2,3[0000-0003-3103-5858]}

¹ Faculty of Engineering, North-West University, South Africa

² Centre for Artificial Intelligence Research, South Africa

³ National Institute for Theoretical and Computational Sciences, South Africa

aldringorima@gmail.com

albert.helberg@nwu.ac.za

marelie.davel@nwu.ac.za

Abstract. Precision agriculture requires the estimation of plant growth stages in real-time. When the plant growth stage is known, the wastage of resources in cultivation, such as nutrients and water, is reduced as only the required resources need to be supplied. Plants at different growth stages, however, have similar morphological features, which can make autonomous growth stage estimation difficult. This paper presents two feature extraction methods for growth stage estimation: one that uses a bank of Gabor filters and morphological operations, and the other that uses pre-trained convolutional neural networks (CNNs) and transfer learning. We test these methods on a publicly available plant growth stage dataset (“bccr-segset”) for two species, canola and radish, grown and captured under indoor conditions. The two proposed feature extraction methods are compared, using support vector machines and boosted trees as classifiers. We find that both methods are suitable for real-time applications, and that CNN features outperform the hand-crafted features, both with regard to speed and accuracy. The best system (VGG-19 features, classified with a radial basis function support vector machine) obtained an accuracy of 98.4% for both species, processing an image in 0.08 seconds.

Keywords: growth stage estimation · precision agriculture · transfer learning · convolutional neural networks · Gabor filters · morphological operations · support vector machines.

1 Introduction

Indoor farming systems, such as vertical farming, are well known for their ability to produce large yields while using fewer resources and taking up less space [26]. In addition, they promote clean food production, as the use of chemicals such as herbicides is not typically required, since farming is done indoors and pests can often be controlled manually. Reducing the use of herbicides also reduces the number of farming resources required. Indoor farming also enables extensive data collection on plants, which is challenging to achieve in conventional farming.

Exploring the impact of different loss functions for anomaly prediction in a mineral processing plant

Morne C. Du Plessis^{1,2[0009–0003–9688–935X]} and Deshendran Moodley^{1,2[0000–0002–4340–9178]}

¹ Department of Computer Science, University of Cape Town,Cape Town, South Africa

duplmor006@myuct.ac.za

² Centre for Artificial Intelligence Research, Cape Town, South Africa
deshen@cs.uct.ac.za

Abstract. Deep neural networks (DNN) have a high potential for predicting and mitigating equipment failure in large industrial plants. However, plant anomalies are rare events that result in extremely unbalanced datasets, which poses a challenge for traditional DNN classifiers. Weighted loss functions such as focal loss and weighted binary cross-entropy (WBCE) have emerged as a promising approach to deal with class imbalance, where higher weightings are assigned to the anomaly class during training. This study proposes three new weighted loss function variants, i.e. weighted polynomial binary cross entropy (WPBCE) loss, weighted hinge loss and weighted squared hinge loss, and systematically evaluates these across three DNN architectures: long short-term memory (LSTM), temporal convolutional network (TCN), and multi-layer perceptron (MLP). The results show that the weighted loss function variants improve recall and yield more stable configurations across all algorithms, when compared to focal loss and WBCE, for predicting the onset of abnormal operating events in a real-world South African mineral grinding mill. Importantly, this work demonstrates that the weighted squared hinge and WPBCE, when combined with the LSTM model, offer a reliable solution for early and accurate anomaly prediction.

Keywords: Anomaly prediction · Loss function · Deep neural networks · Temporal modeling · Class imbalance · Mineral processing · Predictive maintenance.

1 Introduction

Anomaly prediction plays a critical role in industrial mineral processing. The early identification of abnormal operating conditions in grinding mills can prevent costly equipment failures, ensure product quality, and improve overall process efficiency. Recent advances in deep neural networks (DNNs) have shown promise in capturing the complex temporal and nonlinear dynamics inherent in

Modifying class distributions to improve the classification of minority group examples in a class-imbalanced dataset

Banele Mdluli¹ and Terence L. van Zyl²

University of Johannesburg, Johannesburg, JHB, RSA

Abstract. Class-imbalanced datasets are a common occurrence in real-world applications. The imbalance between minority and majority classes exists due to the over-representation of one class compared to another in a dataset. The class imbalance might reflect a system's behaviour over time. However, the class imbalance causes sub-optimal performance for machine learning models that predict the system's future behaviour. Various techniques are used to reduce the negative impact of class-imbalanced datasets on machine learning models. Data resampling techniques are one of the main techniques, and the subdivisions of data resampling techniques include oversampling and undersampling. Oversampling techniques have outperformed undersampling techniques in most studies, and most data resampling techniques are derived from oversampling. However, some oversampling techniques are ineffective when used on minority-class datasets that lack within-class variation and have a high-class imbalance. In this study, an analysis was performed to understand the changes in within-class variation before and after oversampling for nine datasets. Additionally, classification performance was measured for standard and hybrid oversampled datasets. A novel hybrid oversampling technique that uses k-Means and ADASYN was implemented. Hybrid oversampling techniques generated synthetic examples that marginally changed the within-class variation and had the highest F1 score compared to standard oversampling techniques across nine datasets.

Keywords: Class imbalance · Oversampling · ADASYN · Classification algorithm · Within-class variation.

1 Introduction

A class-imbalanced dataset occurs when one class (the majority class) has significantly more examples than the other (the minority class) [22]. Such an imbalance may cause model prediction errors because a model trained on imbalanced data tends to misclassify examples belonging to the minority class [17]. Certain applications of classification models focus on predicting the minority class outcomes (e.g. machine failure, credit fraud detection, early cancer detection) [13, 4, 28]. In those scenarios, class-imbalanced datasets pose a significant threat to model minority class prediction.

Evaluating OCR performance on food packaging labels in South Africa

Mayimunah Nagayi^{1[0009-0009-9241-3059]}, Alice Khan^{2[0000-0003-1011-3693]}, Tamryn Frank^{2[0000-0002-5180-9171]}, Rina Swart^{3[0000-0002-7786-3117]}, and Clement Nyirenda^{1,4[0000-0002-4181-0478]}

¹ Department of Computer Science, University of the Western Cape, Robert Sobukwe Rd, Bellville, Cape Town, 7535, South Africa, 4163113@myuwc.ac.za

² School of Public Health, University of the Western Cape, Robert Sobukwe Rd, Bellville, Cape Town, 7535, South Africa, askhan@uwc.ac.za, tfrank@uwc.ac.za

³ Department of Dietetics and Nutrition, University of the Western Cape, Robert Sobukwe Rd, Bellville, Cape Town, 7535, South Africa, rswart@uwc.ac.za

⁴ eResearch Office, University of the Western Cape, Robert Sobukwe Rd, Bellville, Cape Town, 7535, South Africa, cnyirenda@uwc.ac.za

Abstract. This study evaluates four open-source Optical Character Recognition (OCR) systems which are Tesseract, EasyOCR, PaddleOCR, and TrOCR on real world food packaging images. The aim is to assess their ability to extract ingredient lists and nutrition facts panels. Accurate OCR for packaging is important for compliance and nutrition monitoring but is challenging due to multilingual text, dense layouts, varied fonts, glare, and curved surfaces. A dataset of 231 products (1,628 images) was processed by all four models to assess speed and coverage, and a ground truth subset of 113 images (60 products) was created for accuracy evaluation. Metrics include Character Error Rate (CER), Word Error Rate (WER), BLEU, ROUGE-L, F1, coverage, and execution time. On the ground truth subset, Tesseract achieved the lowest CER (0.912) and the highest BLEU (0.245). EasyOCR provided a good balance between accuracy and multilingual support. PaddleOCR achieved near complete coverage but was slower because it ran on CPU only due to GPU incompatibility, and TrOCR produced the weakest results despite GPU acceleration. These results provide a packaging-specific benchmark, establish a baseline, and highlight directions for layout-aware methods and text localization.

Keywords: OCR · Food Packaging · Ingredient Lists · Nutrition Facts Panels · Tesseract · EasyOCR · PaddleOCR · TrOCR.

1 Introduction

Consumers rely on printed information on food packaging such as ingredient lists, allergen warnings, and nutrition facts to make informed dietary decisions. While regulations require this information to be accurate and legible, packaging often

Investigating the Impact of Multilingual Pre-trained Speech Models on Gender Bias in ASR for Low Resource African Languages

Claytone Sikasote^{1,2[0009–0004–1372–4662]}, Hussein Suleman^{1[0000–0002–4196–1444]}, and Jan Buys^{1[0000–0003–1994–5832]}

¹ University of Cape Town, Cape Town, South Africa
`skscsla001@myuct.ac.za, hussein@cs.uct.ac.za, jan.buys@uct.ac.za`

² University of Zambia, Lusaka, Zambia
`claytone.sikasote@cs.unza.zm`

Abstract. While fine-tuning transformer-based pre-trained speech models improves speech recognition for low resource languages, the approach increases the risk of speaker attribute bias in the resulting target language automatic speech recognition (ASR) systems. This work investigates gender bias in two state-of-the-art pre-trained speech models, MMS and Whisper, fine-tuned for ASR on three African languages: Bemba, Nyanja, and Swahili. We fine-tune models on gender-specific as well as gender-balanced datasets, and estimate and compare gender bias across different settings. Our results show varying degrees of gender bias in the fine-tuned models, even with gender-balanced fine-tuning, suggesting influence from pre-trained models. Inconsistencies in gender-specific fine-tuning further confirm the transfer of bias from pre-trained models. Additionally, an ablation study shows no relationship between training data size and gender bias.

Keywords: Gender Bias · Automatic Speech Recognition · Low Resource Languages · African languages.

1 Introduction

Fine-tuning transformer-based multilingual pre-trained speech models has become a dominant approach to the development of automatic speech recognition (ASR) systems for low-resource languages (LRLs) [5, 34, 33]. Typically, these models are pre-trained on large quantities of multilingual speech data (labeled or unlabeled) and later fine-tuned on a relatively small labeled target language dataset for a downstream speech task. While this approach has led to improved model performance for LRLs, it increases the risk of introducing speaker attribute-specific bias [12], which is the disparity in model performance between different speaker attribute subgroups, for example between male and female speakers.

ASR systems fine-tuned from pre-trained speech models have been reported to exhibit gender bias [4], performing better on male speech in some studies [16,

A Two Stage Pipeline for Automated Caries Detection on Single Tooth Images from Panoramic Radiographs

Christopher J. Hansen^{1[0009-0001-8978-036X]}, Paula Kloehn², Anna-Louisa Kollster², Toni Gehrmann^{2[0009-0007-3543-7420]}, Jonas Conrad^{2[0000-0003-2516-3351]}, Niklas Christoph Koser^{1[0009-0006-0599-2901]}, Christian Graetz^{2[0000-0002-8316-0565]}, Christof Dörfer², Claus-C. Glüer¹, Jan-Bernd Hövener^{1[0000-0001-7255-7252]}, and Coenraad Mouton^{1[0000-0001-8610-2478]}

¹ Section Biomedical Imaging, Dept. of Radiology and Neuroradiology, University Medical Center Schleswig-Holstein (UKSH), Campus Kiel

² Clinic of Conservative Dentistry and Periodontology, University of Kiel, Kiel, Germany

Abstract. Panoramic dental radiographs (OPG) are the only imaging modality that captures the entire dentition in a single exposure. To support dentists with diagnosing caries it is essential to find indications for cavities on those images. While recent deep learning methods show strong results on in-distribution test sets, the generalization on out-of-distribution datasets is mostly untested. In this study, we suggest a two-stage deep learning pipeline for caries detection on single-tooth images extracted from OPGs: (1) image-level classification using a DINO-based transformer backbone and (2) instance-level segmentation using Mask-R-CNN. We perform experiments on data from the University Medical Center Schleswig-Holstein (UKSH). To study generalization, we test models on an out-of-distribution set from the Federal University of Bahia (UFBA) and also evaluate a mixed-domain setting including both UKSH and UFBA data. Further we investigate the influence of strong augmentation techniques. Results show that classification performance is high on in-distribution data but significantly drops when applied to out-of-distribution samples. Segmentation performance is moderate across all settings, with limited robustness under domain shift. These findings suggest that in-distribution results overestimate real-world performance and underscore the importance of evaluating domain shifts in dental AI pipelines.

Keywords: caries · segmentation · classification

1 Introduction

Dental caries is a progressive demineralization of tooth structure caused by bacterial activity, leading to the formation of cavities and potential tooth loss if left

Exploring syllable similarity across South African languages through self-supervised speech representation

Johannes Abraham Louw^[0000-0002-8168-7857]

Natural Language Processing Research Group,
Next Generation Enterprises and Institutions,
CSIR,
Pretoria, South Africa
jalouw@csir.co.za
<https://www.csir.co.za>

Abstract. Syllables are fundamental units in speech production and carry prosodic information, but their acoustic and linguistic properties across different language families are not well understood. This study examines syllable discovery approaches across South African languages using algorithmic syllabification and S5-HuBERT, a self-supervised speech representation model that demonstrates emergent syllabic organization. We analyzed speech recordings from eleven languages representing five language families in South Africa using a systematic comparison of rule-based and data-driven syllable discovery methods. We evaluated both approaches using cross-linguistic consistency measures and acoustic quality assessments across speakers.

Our analysis reveals fundamental differences between the two approaches. Algorithmic syllables demonstrate strong language-family clustering with predominantly language-specific units, while S5-HuBERT units show superior cross-linguistic sharing and weaker family effects. Speaker independence analysis across four experimental phases demonstrates that data-driven methods achieve better acoustic consistency, with the fully data-driven approach reaching near-optimal speaker generalization. These results provide empirical guidance for implementing syllable-based semantic units in multilingual text-to-speech systems for resource-scarce languages.

Keywords: Self-supervised learning · Speech representation · Syllable segmentation · S5-HuBERT · Cross-linguistic analysis · Text-to-speech synthesis

1 Introduction

Developing effective text-to-speech systems for resource-scarce languages requires architectures that can leverage limited high-quality data efficiently. Cross-linguistic generalizability enables such systems to benefit from resource-rich languages and to handle the code-switching scenarios common in multilingual environments. This challenge is particularly pronounced in multilingual contexts

Data Pruning: Redundant, Problematic, and Interdependent Samples

Leon Freese^{1,2[0009–0001–6973–7831]} and Marthinus Wilhelmus
Theunissen^{1,2,3[0000–0002–7456–7769]}

¹ Faculty of Engineering, North-West University, South Africa

² Centre for Artificial Intelligence Research, South Africa

³ National Institute for Theoretical and Computational Sciences, South Africa
`{leонfreese5,tiantheunissen}@gmail.com`

Abstract. The performance of deep learning models is affected by not only data quantity but also data quality. Data pruning is a process by which practitioners can reduce the size of a dataset by only keeping the most important training data points, thereby achieving similar test set performance. We empirically investigate two popular data pruning methods under noisy and noiseless conditions and show that these methods fail in the presence of significant label noise. We highlight that the success of data pruning is distinctly affected by three factors: redundancy in the dataset, the presence of problematic samples, and interdependence between samples. We perform a detailed investigation on commonly used benchmark classification datasets and neural network architectures. We find that our observations are consistent across data distributions and training protocols.

Keywords: Data pruning · Label noise · Deep learning

1 Introduction

Deep neural networks have garnered significant attention due to their remarkable performance on large-scale datasets. However, their effectiveness fundamentally depends on the quality and quantity of training data [23]. With the growing size and complexity of datasets, it has become increasingly evident that not all samples contribute equally to model performance. Datasets can contain statistical outliers [18], mislabelled samples [12], out of distribution instances [10], redundancies [15], imbalances [17], or even *adversarial* examples [4]. Defining and characterizing how each of these issues affect a model’s ability to perform constitutes a wide variety of ongoing investigations in the literature.

Data pruning is the process of removing samples from a train set in a way that allows the model to maintain or improve performance. In addition to possible gains in computational efficiency and robustness, it can also be a useful tool for identifying and analysing training samples to understand their impact on generalization. Data pruning experiments typically involve three steps: 1) ranking training samples based on a *score*, 2) removing a subset of increasing

Restoring Neural Network Plasticity for Faster Transfer Learning

Xander Coetzer^{1[0009–0003–5495–5355]}, Arné Schreuder^{1[0000–0002–9061–1141]},
and Anna Sergeevna Bosman^{1[0000–0003–3546–1467]}

Department of Computer Science, University of Pretoria, South Africa
`u20465026@tuks.co.za`, `an.schreuder@up.ac.za`, `anna.bosman@up.ac.za`

Abstract. Transfer learning with models pretrained on ImageNet has become a standard practice in computer vision. Transfer learning refers to fine-tuning pretrained weights of a neural network on a downstream task, typically unrelated to ImageNet. However, pretrained weights can become saturated and may yield insignificant gradients, failing to adapt to the downstream task. This hinders the ability of the model to train effectively, and is commonly referred to as loss of neural plasticity. Loss of plasticity may prevent the model from fully adapting to the target domain, especially when the downstream dataset is atypical in nature. While this issue has been widely explored in continual learning, it remains relatively understudied in the context of transfer learning. In this work, we propose the use of a targeted weight re-initialization strategy to restore neural plasticity prior to fine-tuning. Our experiments show that both convolutional neural networks (CNNs) and vision transformers (ViTs) benefit from this approach, yielding higher test accuracy with faster convergence on several image classification benchmarks. Our method introduces negligible computational overhead and is compatible with common transfer learning pipelines.

Keywords: Transfer Learning · Computer Vision · Plasticity

1 Introduction

Transfer learning has become a cornerstone of modern computer vision workflows, largely due to the widespread availability of powerful deep learning models pretrained on large-scale datasets such as ImageNet [8]. In typical scenarios, these pretrained models are fine-tuned on a downstream dataset that often differs substantially in size, structure, or domain compared to the dataset used for pretraining. While the transfer learning approach offers significant gains in both performance and data efficiency, it also introduces critical challenges that remain underexplored.

One such challenge is the loss of neural plasticity, i.e., the reduced ability of a neural network to adapt to new data distributions during fine-tuning. This issue arises when certain pretrained weights, particularly those with near-zero magnitudes, result in negligible gradients during backpropagation [14]. Such weights

AfroXLMR-Comet: Multilingual Knowledge Distillation with Attention Matching for Low-Resource Languages

Joshua Sakthivel Raju^{1[0009-0005-4414-2582]}, Sanjay Somasundaram^{1[0009-0007-4172-1815]}, Jaskaran Singh Walia^{1[0000-0002-9255-5446]}, Srinivas R^{1[0009-0000-3669-4727]}, and Vukosi Marivate^{2,3[0000-0002-6731-6267]}

¹ School of Computer Science and Engineering, Vellore Institute of Technology, Chennai, India

joshua.raju2604@gmail.com

² Data Science for Social Impact & African Institute for Data Science and Artificial Intelligence, University of Pretoria, South Africa

³ Lelapa AI

Abstract. Language model compression through knowledge distillation has emerged as a promising approach for deploying large language models in resource-constrained environments. However, existing methods often struggle to maintain performance when distilling multilingual models, especially for low-resource languages. In this paper, we present a novel hybrid distillation approach that combines traditional knowledge distillation with a simplified attention matching mechanism, specifically designed for multilingual contexts. Our method introduces an extremely compact student model architecture, significantly smaller than conventional multilingual models. We evaluate our approach on five African languages: Kinyarwanda, Swahili, Hausa, Igbo, and Yoruba. The distilled student model—*AfroXLMR-Comet*—successfully captures both the output distribution and internal attention patterns of a larger teacher model (*AfroXLMR-Large*) while reducing the model size by over 85%. Experimental results demonstrate that our hybrid approach achieves competitive performance compared to the teacher model, maintaining an accuracy within 85% of the original model’s performance while requiring substantially fewer computational resources. Our work provides a practical framework for deploying efficient multilingual models in resource-constrained environments, particularly benefiting applications involving African languages.

Keywords: Knowledge distillation · Parameter-efficient-training · Data-efficient training · NLP in resource-constrained settings.

1 Introduction

Large language models (LLMs) have become a pillar of modern Natural Language Processing (NLP), achieving state-of-the-art results across various tasks

Code-Switch Pretraining for Improved Cross-Lingual Alignment in Low-Resource Languages

Ruan Visser^{1*}, Trienko Grobler¹, and Marcel Dunaiski¹

Department of Computer Science, Stellenbosch University, Stellenbosch, South Africa
`ruanvisser101@gmail.com, tlgrobler@sun.ac.za, marceldunaiski@sun.ac.za`

Abstract. Cross-lingual language models enable the transfer of linguistic knowledge across languages, however, they often perform worse for low-resource or typologically distant languages. Prior work has explored alignment and adapter methods, but the use of code-switching remains limited and typically confined to fine-tuning with static word substitutions. In this work, we propose an approach that integrates code-switching directly into masked language model pretraining. Instead of applying word substitutions after pretraining, we introduce a multiview probabilistic translation strategy that samples candidate translations based on alignment likelihoods, applying substitutions only to unmasked tokens. This exposes the model to cross-lingual ambiguity and encourages more robust cross-lingual representations. Our results on a diverse set of eight language pairs show that this approach improves zero-shot cross-lingual natural language understanding performance across all languages relative to bilingual baselines. We further observe gains on downstream named entity recognition tasks in most languages when incorporating our code-switched pretraining approach.

1 Introduction

Cross-lingual understanding enables language models to transfer knowledge across languages, supporting scalable deployment in diverse linguistic contexts without relying on large pretraining datasets or expensive labeled downstream tasks [12,2,9]. However, target language performance on cross-lingual tasks still lags behind source language performance, particularly for low-resource languages and those that are typologically distant from source languages. This performance gap provides opportunities for more effective methods to improve cross-lingual alignment.

Many earlier cross-lingual solutions focus on creating language-agnostic models designed to generalize across diverse languages. For example, Artetxe et al. [2] introduced the LASER (Language-Agnostic SEntence Representations) model, which produces sentence embeddings that generalize across 93 languages without additional training. Similarly, Pfeiffer et al. [18] proposed Multiple ADapters

* Corresponding author

Identification of Social Media Users that Perpetuate Xenophobic Attitudes and Hate Speech Narratives in South Africa

Carl du Plessis¹, Michael du Plessis¹, Koen Ronny Mabokela², Abiodun Modupe^{1,3}, Vukosi Marivate^{1,3,4}

¹ Dept. Computer Science, University of Pretoria

² Applied Information Systems, University of Johannesburg

³ Data Science for Social Impact

⁴ Lelapa AI

u88534953@tuks.co.za

Abstract. Social media—particularly X (formerly Twitter)—has become a critical platform for political discourse. It shapes public opinion, influences voter behaviour, and provides real-time insight into contentious issues. Xenophobia, defined as the hostility, or hatred towards foreigners, is a polarising topic in South Africa, especially during election seasons. This paper analyses South African Twitter data from the 2016 and 2021 municipal elections, as well as the 2019 and 2024 national elections, with a focus on Xenophobia-related discourse. We develop a novel machine learning model to identify xenophobic tweets despite the removal of explicit hate speech by platform moderation. Using a labelled dataset of xenophobic tweets, we fine-tuned a transformer-based classifier that achieves over 95% F1-score in distinguishing xenophobic content. We then analyse the prevalence of xenophobic narratives over time, the peaks around election dates, and the user accounts most active in propagating xenophobia. Our results reveal thousands of Xenophobic tweets, peaking sharply during election periods, and show that over half of the top 20 xenophobia-spreading accounts appear affiliated with political figures or parties. We discuss implications for social media policy, election integrity, and community cohesion. We also address ethical considerations such as data privacy, anonymisation of users, and bias. This work contributes a framework for identifying harmful election-related discourse and insights for mitigating the impact of xenophobic narratives on social media.

Keywords: Xenophobia · Hate Speech · Social Media · South Africa · Large Language Models · Elections.

1 Introduction

In democratic societies, elections often heighten debates on identity, immigration, and security. In South Africa, xenophobia—hostility toward foreigners—has

HAPI–FedProx: Heterogeneity–Aware Adaptive Proximal Optimization for Federated Learning

Kudzaishe Mhou^{1[0009–0003–2379–7243]} and Makhamisa
Senekane^{1[0000–0002–0122–3076]}

Institute for Intelligent Systems, University of Johannesburg, Johannesburg, South Africa

216074404@student.uj.ac.za
makhamisa.senekane@nithecs.ac.za; makhamisas@uj.ac.za

Abstract. Federated Learning (FL) is a distributed learning paradigm which entails the training of Machine Learning (ML) models across multiple computing devices, while keeping the training data local to the devices. One of the key challenges of FL is heterogeneity of both the computing devices and data. This challenge might ultimately lead to the FL model instability, slow convergence, and performance degradation. This work introduces Adaptive FedProx, a new FedProx algorithm extension that dynamically modifies its proximal regularisation term in response to real-time heterogeneity detection. In order to direct adaptive regularisation, we present the Heterogeneity-Aware Performance Index (HAPI), a metric that measures the difference between local and global models. We uncover an important trade-off through extensive experiments on CIFAR-10 across Independent and Identically Distributed (IID), mild non-IID, and strong non-IID scenarios: Adaptive FedProx exhibits superior robustness to data heterogeneity, despite a 1.6% performance decrease in homogeneous (IID) settings when compared to FedAvg (87.27% vs. 88.87%, $p < 0.001$). When moving from IID to strong non-IID data, Adaptive FedProx shows 3.5% better robustness with a performance drop of only 22.8% versus FedAvg's 26.3%, and it achieves 67.36% accuracy in strong non-IID scenarios compared to FedAvg's 65.35%. These results imply that, at the expense of a minor drop in performance in homogeneous environments, adaptive regularisation techniques can improve federated learning's resistance to heterogeneous data distributions.

Keywords: Federated Learning · Non-IID · FedProx · Adaptive Regularisation · Heterogeneity-Aware Learning · Convolutional Neural Network

1 Introduction

Federated Learning (FL) is a decentralized approach to machine learning that allows models to be trained across multiple devices or servers while keeping data localized. This technique aims to enhance data privacy by ensuring that sensitive information does not leave its original location, thus mitigating the

Fog-Based Deep Learning for Real-Time Cold Chain Temperature Prediction Using IoT Data

Jeremiah Taguta¹[0000-0001-8199-4307], Jean Frederic Isingizwe Nturambirwe²[0000-0002-1794-7343], and Clement Nthambazale Nyirenda^{1,2}[0000-0002-4181-0478]

¹ Department of Computer Science, University of the Western Cape, Robert Sobukwe Rd, Cape Town, 7535, South Africa
tagutaj@gmail.com

² eResearch Office, University of the Western Cape, Robert Sobukwe Rd, Cape Town, 7535, South Africa
fisingizwe@uwc.ac.za; cnyirenda@uwc.ac.za

Abstract. A third of the food produced globally and in South Africa is lost or wasted annually. Fresh fruits and vegetables (FFVs) contribute 44% of South Africa's wastage, with temperature abuse as the main cause due to their high sensitivity and perishability. Real-time cold chain management with predictive analytics is necessary to control parameters and minimise temperature breaks proactively. While Machine Learning (ML) can predict temperature, cloud deployment causes latency, bandwidth demands, and internet dependency, hindering real-time operations. Fog computing mitigates this by localising ML predictions, an area under-explored for the FFV cold chain. This study investigates fog-based Deep Neural Networks for predicting cold room temperatures in FFV cold chains, utilising IoT data from a South African apple cold room laboratory. SimPy, MinMax scaling, 75%/25% train/test splitting and a sequence length of 2 were used. The fog-based LSTM and GRU ensemble outperformed the cloud-based model with R^2 [0.9081, 0.9245] vs [0.5477, 0.5992], MAE [0.9999 °C, 1.6353 °C] vs [4.2983 °C, 4.6029 °C], MSE [10.0084, 12.0685] vs [50.6492, 56.4827], processing time [0.0840, 0.0847] vs [0.3375, 0.3378], and Pearson R [0.9533, 0.9618] vs [0.7670, 0.7930] (95% confidence). Paired t-test and Wilcoxon test confirm fog's significant superiority. Both had high data utilisation of 99.8%, guaranteeing analysis for transmitted data. High accuracy and low processing time make fog ideal for real-time cold chains, reducing FFV wastage while improving sustainability, affordability, profitability, and food security. Future work will add sensors, fog nodes, compare models, consider different datasets and implement asynchronous sensor fusion and temperature break and cause predictions.

Keywords: Fog Computing · Temperature Prediction · Deep Neural Network · Machine Learning · Cold Chain · Food Wastage · Fresh Fruits and Vegetables

Robust Driver Distraction Recognition via Lightweight Body-Part Association and Object Context on NVIDIA Jetson

Frank Zandamela^{1[0000-0003-2201-1985]}, Mamodike Sadiki^{1[0000-0002-4984-2992]},
Patrick Malatjie^{1[0009-0003-2035-1115]}, Teboho Sekopa^{1[0009-0000-3684-1993]},
and Moloko Manthata^{1[0009-0002-5379-7263]}

Council for Scientific and Industrial Research, Defence and Security, Pretoria 0001,

South Africa

(fzandamela, msadiki, pmalatjie, tsekopa, mmanthata)@csir.co.za

<https://www.csir.co.za/defence-and-security>

Abstract. Driver distraction remains a significant contributor to road accidents; however, existing deep-learning detectors either sacrifice accuracy for speed on resource-constrained hardware or lose generality when confronted with unseen data. This work presents an end-to-end single-stage model that, in one forward pass, jointly identifies the driver, links the driver’s body parts, recognises nearby in-cabin objects, and determines whether the driver is distracted. By embedding spatial and semantic relationships directly into its output vector, the model avoids slow post-processing and significantly reduces false alarms caused by objects that merely appear near the driver. Evaluated on five public datasets and an additional real-world collection that were not used for training, the proposed detector boosts mean F1-score by 0.11 ($\approx 20\%$ relative) over a lightweight baseline while maintaining 39 frames per second on an NVIDIA Jetson Xavier edge device—more than three times faster than a comparable two-stage pipeline. These results demonstrate a driver-distraction detector that simultaneously achieves cross-dataset robustness, real-time performance, and efficient deployment on low-power hardware.

Keywords: Distracted driver detection · Single-stage detector · Edge AI · Body-part association · Real-time inference

1 Introduction

Accurately identifying driver distraction is still a key challenge, and researchers have proposed various deep-learning methods. However, despite recent advancements, there remains a research gap in developing methods with consistent performances across different datasets not used for training and also achieve real-time inference speed on edge devices. There is research focused on developing lightweight distracted driver detection (DDD); for example, Zhao et al. [5] recently proposed a lightweight method that leverages the power of a convolutional neural network and a vision transformer (CoViT). CoViT is a hybrid

Nearest-Class Mean and Logits Agreement for Wildlife Open-Set Recognition

Jiahao Huo¹, Mufhumudzi Muthivhi¹, Terence L. van Zyl¹, and Fredrik Gustafsson²

¹ University of Johannesburg, South Africa `216045414@student.uj.ac.za`,
`mmuthivhi@uj.ac.za`, `tvanzyl@uj.ac.za`

² Linköping University, Sweden `fredrik.gustafsson@liu.se`

Abstract. Current state-of-the-art Wildlife classification models are trained under the closed world setting. When exposed to unknown classes, they remain overconfident in their predictions. Open-set Recognition (OSR) aims to classify known classes while rejecting unknown samples. Several OSR methods have been proposed to model the closed-set distribution by observing the feature, logit, or softmax probability space. A significant drawback of many existing approaches is the requirement to retrain the pre-trained classification model with the OSR-specific strategy. This study contributes a post-processing OSR method that measures the agreement between the models' features and predicted logits. We propose a probability distribution based on an input's distance to its Nearest Class Mean (NCM). The NCM-based distribution is then compared with the softmax probabilities from the logit space to measure agreement between the NCM and the classification head. Our proposed strategy ranks within the top three on two evaluated datasets, showing consistent performance across the two datasets. In contrast, current state-of-the-art methods excel on a single dataset. We achieve an AUROC of 93.41 and 95.35 for African and Swedish animals. The code will be released publicly upon acceptance of this paper.

Keywords: Open-set-recognition · out-of-distribution · wildlife · classification · computer vision · machine learning

1 Introduction

Wildlife classification models have proven to be useful in wildlife monitoring and ecological studies [5]. Several large-scale wildlife classification models have achieved remarkable success over a large variety of animal classes [3, 12, 28, 29, 31, 34]. The largest of which, SpeciesNet, can classify up to 2000 animals.

However, these models are trained under the closed world setting [36]. They perform well over the classes they have seen during training, but will misclassify unknown classes as known classes. Researchers would have to train the model on every species in a region, ensuring that all possible classes have been seen.

Semi-Supervised Object Segmentation via Active Learning for Efficient Ecological Monitoring

Dane Brown¹[0000-0001-7395-7370] and Karen
Bradshaw¹[0000-0003-3979-5675]

Rhodes University, Grahamstown, Makhanda, 6140, South Africa

, d.brown@ru.ac.za

<https://www.ru.ac.za/computerscience/people/academicstaff/profdanebrown/>

Abstract. Deep learning models for instance segmentation have achieved remarkable success, yet their deployment in specialised domains like ecological monitoring is constrained by the prohibitive cost of acquiring high-quality polygonal annotations. This annotation dependency creates a fundamental bottleneck, limiting both scalability and adaptability of these models in real-world conservation scenarios. This paper introduces a data-centric workflow that addresses this challenge through an iterative, multi-stage active learning strategy enhanced with foundation models. The methodology integrates CLIP diversity sampling as an acquisition function with a semi-automated annotation pipeline that combines YOLOv8x6 detection proposals, human-in-the-loop verification, and SAM2-prompted segmentation refinement. A progressive training strategy using YOLOv11s-seg with quality-controlled pseudo-labelling iteratively expands the training dataset while maintaining annotation quality standards. Validated on African Penguin monitoring using Open Images V7 data and independent SANParks field data, experimental results demonstrate that CLIP diversity sampling achieves mAP₅₀ of 0.82 with 400 training samples (without SAM2 refinement), compared to 0.69 for random sampling; with SAM2-refined annotations, performance reaches 0.88 mAP₅₀ using the same 400 samples. Cross-domain generalisation on independent SANParks field data achieves 0.81–0.84 mAP₅₀. The framework reduces annotation requirements while providing a practical solution for deploying instance segmentation in data-scarce domains. Active Learning Instance Segmentation Contrastive LanguageImage Pre-training (CLIP) Ecological Monitoring Segment Anything 2 (SAM2) You Only Look Once (YOLO)

Keywords: Active Learning · Instance Segmentation · Contrastive LanguageImage Pre-training (CLIP) · Ecological Monitoring · Segment Anything 2 (SAM2) · You Only Look Once (YOLO)

Scaling behavior of Encoder Language Models in Low-Resource Settings

Ruan Visser^{1*}, Trienko Grobler¹, and Marcel Dunaiski¹

Department of Computer Science, Stellenbosch University, Stellenbosch, South Africa
`ruanvisser101@gmail.com, tlgrobler@sun.ac.za, marcel.dunaiski@sun.ac.za`

Abstract. Pretraining language models for low-resource languages poses significant challenges due to scarce and poor-quality data, a lack of comprehensive evaluation benchmarks, and often limited computational resources. Research on compute-optimal language modeling typically focuses on scaling up decoder language models efficiently for high-resource languages. While some studies have investigated the down-scaling of encoder language models for low-resource languages, they often prioritize optimizing for computational constraints rather than pretraining text volume constraints. We address this research gap by analyzing the scaling behaviors of encoder language models which use the Replace Token Detection (RTD) and Masked Language Modeling (MLM) objectives under limited pretraining text volumes.

By downsampling three different high-resource languages (English, French, Korean) and two low-resource languages (Xhosa and Swahili), we simulate varying degrees of data scarcity and evaluate downstream performance using established benchmarks such as the GLUE benchmark for English, FLUE for French, KLUE for Korean, and MasakhaNEWS for Xhosa and Swahili. Our findings demonstrate that optimal MLM accuracy scales logarithmically with increasing pretraining text volume across these diverse languages. Additionally, our results show that RTD models consistently outperform MLM models in low-resource scenarios, achieving superior downstream performance with pretraining text volumes smaller than 1000MB for downsampled high-resource languages. However, we find that RTD performs worse than MLM for Xhosa and Swahili. We also find that dynamic masking significantly improves MLM accuracy in these settings. Furthermore, our results show that smaller models are more effective for smaller pretraining text volumes, highlighting the importance of adjusting model size according to data availability in order to maximize performance and efficiency.

1 Introduction

Recent work on compute optimal language modeling has enabled researchers to train better performing large language models (LLMs) while minimizing computational requirements. Notable studies such as Hoffmann et al. [11] and Kaplan

* Corresponding author

Enhancing Target Re-Identification via Model Fusion and Knowledge Distillation of Pre-trained Foundation Models

Tendai Shoko¹[0000-0002-1245-4267] and Terence L. van Zyl^{1,2}[0000-0003-4281-630X]

¹ Institute for Intelligent Systems, University of Johannesburg, Johannesburg, South Africa
223029866@student.uj.ac.za

² CAIR, Institute for Intelligent Systems, University of Johannesburg, Johannesburg, South Africa
tvanzyl@uj.ac.za

Abstract. Target re-identification (re-ID) systems face critical deployment challenges balancing accuracy with computational efficiency in resource-constrained environments. This paper presents a novel framework integrating Mixture-of-Experts (MoE) with Knowledge Distillation (KD) to leverage pre-trained foundation models effectively. The framework employs dynamic expert selection to combine CLIP and ALIGN models, then distills their collective knowledge into a compact student architecture. Experimental evaluation on VeRi-776 and Market-1501 demonstrates 75.2% and 76.1% mAP respectively, while reducing inference time by 50% and model parameters by 94% compared to the MoE ensemble (and approximately 92% vs CLIP fine-tuning). Comprehensive ablation studies validate the synergistic benefits of MoE and KD components, showing improved cross-domain performance with 12.9% mAP degradation versus 15.3% for conventional methods. The results demonstrate MoE-KD as a practical solution for real-world re-ID deployment.

Keywords: target re-identification · mixture-of-experts · knowledge distillation · foundation models · computational efficiency · computer vision

1 Introduction

Real-world surveillance systems process millions of images daily across hundreds of cameras, requiring re-identification models that achieve high accuracy while maintaining computational efficiency for real-time operation. A typical urban surveillance network with 200 cameras generating 30 frames per second must process 6,000 images per second. Existing foundation models like CLIP [2] achieve impressive accuracy but require 150ms inference time per image, making deployment in such scenarios computationally infeasible. This computational bottleneck limits the adoption of state-of-the-art models in practical applications where response time and resource constraints are critical.

Heart Disease Prediction: A Comparative Study of Optimizers' Performance in Deep Neural Networks

Chisom Chibuike^{1[0009–0005–5589–814X]*} and Adeyinka
Ogunsanya^{2[0000–0002–4739–5551]}

¹ University of Nigeria, Nigeria.

`chisom.chibuike.246093@unn.edu.ng`

² Sydani Institute for Research and Innovation, Sydani Group, Nigeria.

`adeyinka.ogunsanya@sydani.org`

Abstract. Optimization has been an important factor and topic of interest in training deep learning models, yet less attention has been given to how we select the optimizers we use to train these models. Hence, there is need to dive deeper into how we select the optimizers we use for training and the metrics that determine this selection. In this work, we compare the performance of 10 different optimizers in training a simple Multi-layer Perceptron model using a heart disease dataset from Kaggle. We set up a consistent training paradigm and evaluate the optimizers based on metrics such as convergence speed and stability. We also include some other Machine Learning Evaluation metrics such as AUC, Precision, and Recall, which are central metrics to classification problems. Our results show that there are trade-offs between convergence speed and stability, as optimizers like Adagrad and Adadelta, which are more stable, took longer time to converge. Across all our metrics, we choose RMSProp to be the most effective optimizer for this heart disease prediction task because it offered a balanced performance across key metrics. It achieved a precision of 0.765, recall of 0.827, and an AUC of 0.841, along with faster training time. However, it was not the most stable. We recommend that in less compute constrained environment, this method of choosing optimizers through a thorough evaluation should be adopted to increase the scientific nature and performance in training deep learning models.

Keywords: Deep Neural Network · Optimizer · Stability

1 Introduction

In recent years, Deep Neural Networks (DNNs) have been central to advances in artificial intelligence, driving progress across a wide array of applications, including autonomous systems, medical diagnostics, natural language processing, and

* This work was done while Chisom Chibuike was a research intern at SAIL Innovation Lab, supervised by Adeyinka Ogunsanya.

Part II

Volume I: Socio-technical and Human-Centred AI

AI-Enabled Cybersecurity Implementation: A Case Study of Critical Success Factors in a South African State-Owned Entity

Awonke Mamane [0009-0006-1530-0606] and Rennie Naidoo [0000-0001-8392-1136]

Department of Information Systems, University of the Witwatersrand, Johannesburg, South Africa

2138857@students.wits.ac.za, Rennie.Naidoo@wits.ac.za

Abstract. The integration of AI into cybersecurity is essential for addressing complex and evolving threats. However, much of the existing AI implementation research emphasises either technical or social dimensions, neglecting their socio-technical interdependence. This study addresses this gap by identifying the critical success factors (CSFs) for AI-driven cybersecurity implementation through a socio-technical lens. Using an interpretive case study of a South African state-owned entity, the research draws on thematic analysis of in-depth interviews with technical staff and end-users. Findings reveal that successful implementation depends on the interplay between technical and social elements. Key technical CSFs include data quality, scalability, automation, and efficiency, while social CSFs encompass change acceptance, top management support, user awareness, ethical considerations, human oversight, and usability. Crucially, the study confirms that neither technical nor social factors alone are sufficient and that effective implementation depends on their interdependence. By applying a socio-technical perspective, the research offers a more balanced understanding of AI-driven cybersecurity and presents a framework to support practitioners in implementing socially integrated, technically robust solutions. Future research should further examine how human-AI collaboration can be socio-technically integrated to enhance trust, ensure ethical compliance, and improve the operational reliability of AI-enabled cybersecurity systems within organisational settings.

Keywords: artificial intelligence, cybersecurity, socio-technical systems, case study, critical success factors, implementation.

RAG Evaluation: From Model-Centric Benchmarks to System-Level Metrics

Alta de Waal¹ , Daniel van Niekerk¹ , Florian Donhauser² ,
Salmaan Suliman¹ , and Dehan Lamprecht³ 

¹ BMW IT Hub, South Africa

² BMW Group, Munich, Germany

³ Department of Applied Mathematics, University of Stellenbosch, South Africa

Abstract. As Large Language Models (LLMs) increasingly support and automate business-critical workflows, the need for robust evaluation frameworks becomes paramount. This paper proposes a system-level testing approach designed to assess the performance and reliability of LLM-based applications integrated into enterprise processes. Moving beyond model-centric benchmarks, the framework adopts principles from software engineering, including black-box and end-to-end testing, to evaluate real-world outcomes in retrieval-augmented generation (RAG) systems. It features modular performance indicators such as context precision, hallucination detection, business tonality alignment, and answer correctness, many harnessing the LLM-as-a-Judge methodology. Answer correctness is used as a case study for the design of interpretable performance indicators grounded in concepts from information retrieval while considering the objectives of business and technical stakeholders. Empirical evaluations in four use cases demonstrate how this approach enables organisations to validate not only the accuracy of their systems but also business relevance.

Keywords: retrieval augmented generation · evaluation · answer correctness.

1 Introduction

Large Language Models (LLMs) have become a core technology powering modern AI systems, enabling everything from conversational assistants to code-generation tools. Robust evaluation is essential to ensure that these models perform reliably, safely, and ethically in real-world applications, where failures can result in misinformation, biased output, or significant business and reputational risks [5,7]. Traditionally, evaluation efforts have focused on benchmarking LLMs as standalone entities, using model-centric metrics such as perplexity, BLEU [21], ROUGE [15], and METEOR [2]. These metrics assess intrinsic linguistic capabilities, such as text generation, comprehension, and summarisation, and provide valuable information to improve foundational models [24,27]. However, as

Exploring Postgraduate Students' Use and Perceptions of Generative AI as a Research Support Tool: A South African Case

Mfowabo Maphosa^{1*}[0000-0003-3702-6821], Lucas Khoza²[0000-0002-7043-6462], Cyril Tlomatsana³[0009-0007-4218-7861] and Winnie Pitjo⁴[0009-0005-5278-3363]

¹ University of Pretoria, Pretoria, South Africa
mfowabo.maphosa@up.ac.za

² University of South Africa, Pretoria, South Africa
khozalt@unisa.ac.za

³ The IIE Varsity College, Johannesburg, South Africa
stlomatsana@varsitycollege.co.za

⁴ FNB, Johannesburg, South Africa
Mashienyane.Pitjo@fnb.co.za

Abstract. The emergence of generative artificial intelligence (GenAI) and its applicability within higher education institutions (HEIs) has gained momentum worldwide. GenAI tools have caused a paradigm shift in education including students' research activities. Despite various studies being conducted on GenAI tools in education, most research remains concentrated on developed countries, with limited attention to how these technologies are perceived in developing nations. Therefore, this study explores the usage and perceptions of GenAI tools among postgraduate students enrolled for postgraduate diplomas, honours and master's degrees at a private HEI in South Africa. Using a mixed-methods approach, the study surveyed 75 students to understand their usage and perceptions of GenAI tools for supporting research activities. The findings reveal that almost three-quarters of the students use GenAI tools, particularly ChatGPT, and have a positive attitude towards the use of GenAI tools to support their research activities. The high usage of GenAI tools is attributed to their capability to generate research ideas, summarise articles, and simplify difficult concepts. Over a quarter of the surveyed students do not use GenAI tools due to concerns about plagiarism, bias, privacy and the potential to impair cognitive development. 78% of the students are familiar with the institution's policy on GenAI. Most students use GenAI tools to generate research topics and ideas. HEIs should assert balance in integrating GenAI tools to support students' research activities. HEIs should further invest in universal frameworks that will serve as guidelines for using GenAI tools in scholarly activities without violating academic and ethical integrity.

Keywords: Generative Artificial Intelligence, Higher Education Institutions, Research Support, Postgraduate Students.

An Awareness Framework for Sustainable Selection of LLMs in Business

Natascha Brughitta Anchia and Andreas Martin [0000-0002-7909-7663]

FHNW University of Applied Sciences and Arts Northwestern Switzerland, School of Business, Rigggenbachstrasse 16, 4600, Olten, Switzerland
natascha.brughittaanchia@outlook.com, andreas.martin@fhnw.ch

Abstract. The environmental impacts of large language models (LLMs) often remain invisible in business adoption. This paper presents an awareness framework to support the sustainable selection of LLMs, developed using a design science research approach within the marketing department of a major European engineering and technology company. Addressing the lack of transparency and emissions data from LLM providers, the artefact calculates electricity use, carbon emissions, and material impacts of inference tasks and visualises them in an interactive dashboard. Evaluation workshops with stakeholders from marketing, sustainability, and AI strategy confirmed the framework's potential to foster awareness, support sustainable decision-making, and align AI use with corporate environmental goals and the UN Sustainable Development Goals (SDG). The framework is transferable to other business contexts.

Keywords: Sustainable Artificial Intelligence · Large Language Models (LLM) · Business AI Adoption · Design Science Research

1 Introduction

The rapid adoption of large language models (LLMs) in business offers efficiency gains but raises growing concerns about environmental sustainability [23,8,29]. While organisations increasingly align with climate goals and frameworks such as the UN Sustainable Development Goals (SDGs), the environmental footprint of LLMs remains largely invisible in decision-making processes [33,7,10,21,12]. High energy use, carbon emissions, and material consumption during inference are often overlooked, due to limited transparency from model providers and a lack of accessible sustainability tools [27].

Although sustainability reporting is gaining traction, businesses still lack mechanisms to evaluate the environmental and financial impacts of LLMs in real-world functions [27,33]. Decisions are often based on performance or novelty, rather than sustainability. The absence of standardised methods to compare general-purpose and task-specific models further hinders responsible selection [5]. This gap is particularly evident in marketing departments, where LLM use is rising but sustainability remains underexplored [8,13,19,34].

Synthesizing Insights on AI in Africa: A Comprehensive Qualitative Analysis

Arti Ramanathan^{1[0009–0008–9891–1270]}, Dongpeng Huang^{1,2[0000–0002–7593–4612]}, and James E. Katz^{1[0000–0002–7441–1078]}

¹ College of Communication, Boston University, Boston, MA 02215, USA
{artiram, dphuang, katz2020}@bu.edu

² The Derek Bok Center for Teaching and Learning,
Harvard University, Cambridge, MA 02138, USA
dongpeng_huang@fas.harvard.edu

Abstract. This exploratory study synthesizes insights from 14 in-depth expert interviews to examine the state of artificial intelligence (AI) in Africa across six thematic areas: infrastructure development, governance frameworks, cultural preservation, linguistic equity, the startup ecosystem, and youth empowerment. Through systematic qualitative analysis, we identify preliminary challenges, opportunities, and context-specific strategies for responsible AI growth on the continent. As an exploratory investigation, the sample size aligns with established guidelines for thematic saturation in qualitative research. Verbatim transcripts were analyzed using Braun and Clarke's thematic analysis framework, with sentiment quantification through AI-assisted coding and manual validation. The initial analysis revealed significant AI readiness gaps, with less than 1% of global AI computing capacity located in Africa. While governance frameworks show promise through initiatives like the African Union's Continental AI Strategy, implementation remains fragmented. Cultural preservation efforts generated the most positive sentiment (52%), demonstrating successful AI applications in heritage digitization. Conversely, infrastructure and policy discussions revealed predominantly negative sentiment, reflecting systemic barriers. Despite challenges, participants highlighted emerging opportunities through public-private partnerships, grassroots language technology innovations, and youth capacity-building programs. These preliminary findings contribute foundational empirical evidence for developing Africa-centric AI strategies that balance technological advancement with cultural preservation and inclusive development, while establishing a framework for future large-scale research across the continent.

Keywords: Artificial Intelligence · Africa · Qualitative Research · Sentiment Analysis · Thematic Analysis · AI Policy · Technology Governance · Linguistic Equity

1 Introduction

The global deployment of artificial intelligence reveals fundamental tensions between universal technological promises and context-specific realities. While

Factors influencing the use of Autonomous Vehicles on Traditional Road Infrastructure in the Western Cape

Naasir Parker^{1,2[0009-0006-7077-0889]} and Walter Ferreira Uys^{1,2[0000-0001-7709-9326]}

¹ Department of Information Systems, University of Cape Town, Cape Town, South Africa

² CITANDA, University of Cape Town, Cape Town, South Africa

Walter.Uys@uct.ac.za

Abstract. This study examines the factors that influence the use of autonomous vehicles (AVs) on traditional road infrastructure in the Western Cape, with a focus on AV compatibility with existing infrastructure. The Diffusion of Innovations Theory serves as a conceptual framework for assessing the potential and challenges of AV implementation. A qualitative approach was employed through semi-structured interviews with experts from the Departments of Mobility, Infrastructure, and Environmental Affairs in the Western Cape. A thematic analysis of the interviews indicates that the Western Cape's paved roads are mostly suitable for AV implementation. However, some adaptation would be necessary in rural areas with gravel roads and limited connectivity. Findings indicate that environmental and economic factors, such as funding limitations and public purchasing preferences, negatively influence AV adoption. Political advantage, however, may positively influence the diffusion process. Surprisingly, the study suggests that AVs may need to adapt to existing road infrastructure rather than vice versa, which contrasts with the established literature on AV implementation in developed regions.

Keywords: Autonomous Vehicles, Road Infrastructure, Diffusion of Innovation.

1 Introduction

Autonomous Vehicles (AVs) have emerged as a disruptive technology in the automotive industry, paralleling the development of Electric Vehicles (EVs), Hybrid Electric Vehicles (HEVs), and Artificial Intelligence (AI) [1–4]. Companies such as General Motors (Cruise), Geely, Waymo, Volvo, Tesla and others are leading this development [5] with level four automation already achieved and level five automation targeted as the next frontier. This research, however, is primarily focused on developed countries, such as the United States, the Netherlands, the United Kingdom, Canada, Australia, Italy, and Russia [3, 6]. The limited research on the adoption of AVs in Southern Africa is mainly focused on driver acceptance of AVs [7]. Several factors appear to influence user acceptance of AVs in the Western Cape [7]. These include improved driving safety, the efficacy of the technology, seamless integration with existing driving systems, and a positive perception among drivers who are familiar with or have experience

Large Language Model Powered Agentic Framework for Cholera Risk Prediction with Explainable ML and Statistical Insights

Paul Jideani^{1,2[0000-0001-5836-6660]} and Aurona Gerber^{1,3,4[0000-0003-1743-8167]}

¹ Department of Computer Science, University of the Western Cape, South Africa

² Boston City Campus, Stellenbosch, South Africa

³ Center for AI Research (CAIR), South Africa

⁴ Department of Information Science, Stellenbosch University, South Africa

pcijideani@gmail.com

Abstract. Interpreting the complex and multifactorial risk factors driving Cholera outbreaks remains a critical challenge for public health, particularly across diverse environmental and socio-economic contexts. This paper presents an integrated agentic framework that combines explainable machine learning (ML), statistical analysis, and a language model-powered question-answering system to support Cholera risk interpretation and public health decision-making. Using a multi-country dataset spanning 2000–2025, the framework applies three interpretable ML models, Explainable Boosting Machines (EBM), Natural Gradient Boosting (NGBoost), and TabNet, to predict Cholera incidence based on environmental, socio-economic, and infrastructural variables. In parallel, statistical methods including Pearson and Spearman correlation, and multivariate linear regression are used to validate and quantify associations between predictors and disease outcomes. A LangChain-powered agent, implemented with LangGraph, is integrated into the system to interpret model outputs, analyse tabular results, and generate expert-like responses to natural language queries. The agent draws evidence from multiple CSV-based analyses, including feature importance scores, correlation matrices, regression coefficients, and model performance comparisons to provide grounded, interpretable answers and policy recommendations. A Streamlit interface enables interactive exploration of Cholera risk factors by researchers, health professionals, and policy stakeholders. Results show strong agreement among models on key predictors, such as rainfall frequency, stagnant water presence, and open defecation, with statistically significant relationships confirmed through regression analysis. The EBM model achieved the lowest RMSE (0.421), indicating superior predictive performance. This work demonstrates how explainable AI and LLM agents can be combined into a transparent, interpretable, and actionable framework for public health analytics, offering valuable insights into data-driven disease prevention strategies.

Keywords: LLM, Decision support, LangChain, Interpretable AI, Statistical analysis

A Scoping Review on Generative AI prompting to optimize the workflow of healthcare professionals in sub-Saharan Africa

Oluwamayowa Ogundaini^{1[0000-0002-3243-6082]} and Lisa-Dionne Morris^{2[0000-0002-6348-7675]}

¹ AFDA School of Business Innovation and Technology, Cape Town, South Africa

² School of Mechanical Engineering, University of Leeds, Leeds, United Kingdom

Abstract. Generative artificial intelligence (AI) applications have enhanced democratization of information to the extent that industry professionals can automate routine tasks, gain insights from complex data and execute tasks more efficiently through generation of text, image, and audio content. Although these applications augment human capabilities, there are concerns about veracity of AI prompting, which results in hallucinations that could have dire consequences on clinical workflow of healthcare professionals. The impact of prompting patterns on optimization of clinical workflows at points-of-care remains nascent with limited evidence especially in healthcare sectors of sub-Saharan Africa. This review explores existing literature on how generative AI prompt engineering optimizes clinical workflow of healthcare professionals by adopting Arksey and O'Malley five-stage scoping review framework to analyze peer-reviewed publications. A comprehensive search strategy was conducted in scholarly databases, including PubMed, IEEE Xplore, and Google Scholar between 2019 and 2025. The study highlights AI prompt engineering strategies, how prompting affects clinical and administrative activities, and how limitations of generative AI prompting could be addressed. Evidence of generative AI prompt engineering are limited in SSA while the Global North and China are the most dominant regions in the discourse. Consultations, clinical decision support, record summaries and documentation, research and prescription recommendations are leading activities in which AI prompting is perceived as most significant. To conclude, this study provides insights for health managers, healthcare professionals, data scientists, ethicists, health IT experts, human-computer interaction practitioners, and researchers on standardizing integration of generative AI use at points-of-care.

Keywords: Generative AI, Prompt Engineering, Large Language Models, Clinical Workflow, Hallucination, Sub-Saharan Africa

Detection Of Power Line Insulator Defects Using YOLOv10-N

Kgampu Shawn Papi¹ and Terence van Zyl¹

¹ Institute for Intelligent Systems, University of Johannesburg, Johannesburg, South Africa
papiks@eskom.co.za, tvanzyl@uj.ca.za

Abstract. Visual inspection remains a common approach for assessing composite insulators, with unmanned aerial vehicles (UAVs) increasingly preferred due to their efficiency and reduced error rates. Recent developments have integrated artificial intelligence (AI) algorithms directly into UAV hardware to enable faster processing; however, such systems require optimized models owing to limited onboard computing resources. The recently introduced YOLOv10-N model, which offers greater efficiency compared to its predecessors, demonstrates potential for detecting insulator defects on resource-constrained UAV platforms. This study evaluates the effectiveness of YOLOv10-N for this application.

Keywords: YOLO, Insulator, Visual inspection, Image, Embedded systems.

1 Introduction

Visual inspection remains one of the most widely used methods for assessing the condition of composite insulators. This technique is generally effective, as many types of surface damage can be readily identified during climbing inspections or, in some cases, from the ground, a helicopter, or an unmanned aerial vehicle (UAV). UAV technology is rapidly becoming the preferred method for aerial insulator inspection because traditional alternatives, such as helicopter-based surveys, are time-consuming, costly, and susceptible to human error [1].

Recent developments indicate a shift toward integrating artificial intelligence (AI) applications directly into UAV hardware or single-board computers mounted on the UAV. This approach offers lower latency compared to UAV systems that stream imagery to cloud-based AI applications for processing [2]. UAVs with onboard AI capabilities are receiving growing research attention in insulator inspection applications [3, 4]. However, single-board computers (e.g., Raspberry Pi) and UAV hardware generally have limited processing power, restricting them to AI models optimized for devices with constrained computational resources.

Panigrahy et al. [4] deployed a YOLOv8-N model on a Raspberry Pi 4 mounted on a UAV and demonstrated the model's ability to detect insulators under varying climatic conditions. Notably, the recently introduced YOLOv10-N model requires 28% fewer parameters than YOLOv8-N for object detection [5], thereby reducing the

Part III

Volume I: Responsible and Ethical AI

Artificially Generated Minorities (AGMs): The Veneer of Algorithmic Bias Correction

Ibifuro Robert Jaja^{1[0000-0003-1101-8591]}

Centre for Philosophy and Artificial Intelligence Research (PAIR), Friedrich-Alexander-Universität Erlangen-Nürnberg, Nürnberger Str. 74, 91052, Erlangen, Germany
Department of Philosophy, Rivers State University, Rivers State, Nigeria

Ibifuro.jaja@fau.de

Abstract. Algorithms often reinforce societal biases and stereotypes. This is especially concerning for minorities, who are disproportionately impacted by it, thereby threatening their further marginalization. Data fundamentalists frame this issue of algorithmic bias as stemming from data bias, indicated by the underrepresentation of some groups (minorities) in the datasets. Consequently, measures adopted to address algorithmic bias have been data-focused. A relatively recent data-focused measure adopted to address this issue is the deployment of what I term artificially generated minorities (AGMs)—synthetic data used to increase the representation of underrepresented groups (minorities) in algorithms' training datasets. Data fundamentalists make two central claims about AGMs, which I term the representation claim, which holds that AGMs are representative of minorities, and the normative intervention claim, which holds that the deployment of AGMs addresses algorithmic bias. In this paper, I argue that AGMs do not meet these claims, particularly in the context of algorithmic recruitment. First, I demonstrate that AGMs do not capture the experience of historic and systemic oppression, which defines minority status. Hence, I contend that they do not meaningfully represent minorities. Second, I demonstrate that while AGMs facilitate the realization of the futuristic component of an adequate normative intervention, they undermine the reparative component. Thus, I contend that AGMs do not adequately address algorithmic bias. Finally, I briefly highlight that the failure of AGMs to meet these claims indicates that a data-focused framing of algorithmic bias is overly simplistic and does not account for all the complexities involved in the issue of algorithmic bias and its correction, particularly in the context of algorithmic recruitment.

Keywords: Algorithmic Bias, Artificially Generated Minorities (AGMs), Data Fundamentalism, Representation Claim, Normative Intervention Claim.

Over-Researched and Under-Resourced: The ELSA Approach to Transdisciplinary AI Research in Low-Trust Neighbourhoods

Nanou van Iersel^{1[0009-0007-0650-4030]}, Majsa Storbeck^{2[0009-0003-4584-0046]}, Marlon Kruizinga^{3[0009-0001-6400-118X]} and Michaël Grauwde⁴

¹ Erasmus University Rotterdam, Burgemeester Oudlaan 50, 3062 PA Rotterdam, the Netherlands
vaniersel@law.eur.nl

² Erasmus University Rotterdam, Burgemeester Oudlaan 50, 3062 PA Rotterdam, the Netherlands

³ Erasmus University Rotterdam, Burgemeester Oudlaan 50, 3062 PA Rotterdam, the Netherlands

⁴ Delft University of Technology, Mekelweg 5, 2628 CD Delft, the Netherlands

Abstract. This paper discusses ELSA (Ethical, Legal, and Social Aspects of technology) as an emerging methodology for transdisciplinary AI research, characterized by anticipatory technology assessment through close collaboration with diverse (societal) stakeholders. We offer a methodological reflection based on a 1,5 year-long case study on public safety and AI in Lombardijen, a neighbourhood in Rotterdam, the Netherlands, where we engaged residents as citizen stakeholders. Lombardijen is paradoxically under-resourced, meaning historically neglected and stigmatized as a ‘problem district’, yet over-researched, i.e. scrutinized by countless researchers who engage in what has been called ‘drive-by’ research – driving by, extracting data, and disappearing, often without benefits for the community. The community’s ensuing alienation from governmental and academic institutions means that citizens’ valuable contextual knowledge is often overlooked in public deliberation on AI. This raises our research question: How can citizens in low-trust neighbourhoods be meaningfully and reciprocally engaged in transdisciplinary AI research, and what does an ELSA approach offer in this regard? The paper details our experiences in Lombardijen respectively from ethical, legal, social, and technological perspectives. We candidly discuss our learnings, (modest) successes and limitations, ultimately emphasizing the importance of situated responsibility as a precondition for transdisciplinary AI research.

Keywords: Transdisciplinary Research, ELSA, Citizen Engagement, Artificial Intelligence, Digital Technologies, Public Safety, Neighbourhoods.

1 Introduction

The history of scientific practice and research is also a history of extractivism: the socio-ecologically destructive appropriation and subjugation of human and natural resources through domination without accountability for (grave) consequences [1]. Smith [2] has powerfully described how Indigenous communities have experienced Western

From Principles to Practice: Extracting Human-Centered AI Components from Governance Frameworks

Cindy van den Berg^{1[0009-0007-0567-242X]} and Hanlie Smuts^{2[0000-0001-7120-7787]}

^{1,2} University of Pretoria, Private Bag X20, Hatfield, Pretoria 0028, South Africa
¹cinvandenberg@gmail.com, ²hanlie.smuts@up.ac.za

Abstract. Artificial intelligence (AI) offers numerous benefits, including increased automation levels, but can also harm individuals and communities, which raises concerns. The field of human-centered AI (HCAI) was created to address human involvement and the consideration of human rights and values involved in the development of AI systems. Document analysis of governance frameworks was used as a research approach to identify components that contribute to human-centered AI solutions. Four authoritative bodies were selected as the principal sources of AI-related principles, standards, and guidelines, based on their recognized authority, global relevance, and comprehensive regulatory scope. Thirty-seven human-centered components were extracted and classified into five classification categories: human-centered values and ethics, user experience and human interaction, data and model governance, technical robustness and system performance, and AI system capabilities and design considerations. The identified components can be used to develop AI solutions that are human-centered and uphold legal integrity, fundamental freedoms, and the principles of democratic governance. Including these components in a formal development methodology can assist in developing AI solutions that are human-centered, free from bias, beneficial to humans, supportive of human capabilities, and aligned with ethical, transparent, reliable, trustworthy, and explainable principles.

Keywords: Human-centered, artificial intelligence, HCAI, human-centered framework.

1 Introduction

Artificial intelligence (AI) is infiltrating all industries and our daily existence [1-3]. It offers many benefits but can also cause harm to people and communities [1, 4-8]. Current AI implementation focuses on algorithmic performance rather than its impact on humans [8-10]. AI enables increased levels of automation, with the potential to replace humans, which raises concerns [1-4, 7, 8, 11-13]. The field of human-centered AI (HCAI) arose to address human involvement and the consideration of human rights and values involved in the development of AI systems [1, 4, 7, 11, 14, 15]. HCAI also focuses on creating AI systems that benefit humans and amplify human expertise and skills [7, 8, 12, 16, 17]. AI solutions should be developed, rolled out, and utilized in a

Part IV

**Volume II: Algorithmic and
Data Driven AI**

MAGE: Multi-Head Attention Guided Embeddings for Low Resource Sentiment Classification

V. Vashisht¹[0009–0002–3466–4676], S. Singh¹[0009–0004–1652–595X], M. Konduskar¹[0009–0009–3203–6976], J.S. Walia¹[0000–0002–9255–5446], and V. Marivate²[0000–0002–6731–6276]

¹ School of Computer Science and Engineering, Vellore Institute of Technology
² University of Pretoria

Abstract. Due to the lack of quality data for low-resource Bantu languages, significant challenges are presented in text classification and other practical implementations. In this paper, we introduce an high-performing model combining Language-Independent Data Augmentation (LiDA) with multi-head attention based weighted representations to selectively enhance critical features within the embedding space and improve text classification performance. This integration allows us to create stable data augmentation strategies that are accurate across various linguistic contexts, ensuring that our model can handle the unique syntactic and semantic features of Bantu languages. This approach not only addresses the data scarcity issue but also sets a foundation for future research in low-resource language processing and classification tasks.

1 Introduction

Text classification is one of the most widely explored tasks in Natural Language Processing (NLP) due to its diverse applications, including spam detection, sentiment analysis, and topic modeling. Despite the impressive advancement achieved through deep learning, these methods rely heavily on large amounts of labeled data, posing a challenge for low-resource languages [13, 14]. African languages in general exemplify this challenge, as the scarcity of annotated datasets limits the development of accurate text classification models [1, 2]. Data augmentation has emerged as a promising solution for addressing data scarcity by generating synthetic data from original datasets [7]. Traditional augmentation techniques, including synonym replacement, sentence back-translation, and generative models, rely heavily on language-specific resources such as pre-trained word embeddings, language models, or linguistic databases, such as WordNet [4, 11, 15, 24]. This language dependence makes these approaches less accurate for underrepresented languages, such as Bantu languages, which lack these linguistic resources [27]. To overcome these limitations, Language-Independent Data Augmentation (LiDA) was introduced [22], operates on sentence embeddings (embedding space) rather

than surface text at the word/token level. LiDA transforms sentence embeddings to generate synthetic data, bypassing the need for language-specific resources. Building upon this foundation, we propose MAGE (Multi-Head Attention Guided Embeddings), a framework designed to enhance text classification performance for low-resource languages. MAGE extends the LiDA framework by introducing significant innovations to the embedding and augmentation process. Specifically, it replaces the traditional Denoising Autoencoder with the Variational Autoencoder (VAE) to enable more expressive and diverse synthetic embeddings. Additionally, MAGE incorporates a novel multi-head attention mechanism that selectively emphasizes salient features in the embeddings. This focus on multi-head attention improves the model’s capacity to capture critical syntactic and semantic nuances, making it particularly accurate for low-resource languages. Using the AfriSenti SemEval dataset [12], a collection of tweets annotated with positive, negative, and neutral sentiments for Kinyarwanda, Swahili, and Xitsonga, we evaluate the performance of MAGE in sentiment classification. Our results demonstrate that MAGE outperforms baseline approaches in low-resource settings. Moreover, comparative analyses highlight the advantages of MAGE over self-attention-based models, further establishing its value as a stable framework for addressing the challenges posed by data scarcity in low-resource languages.

Bantu languages pose particular challenges for natural language processing due to their linguistic complexity. They exhibit rich morphology, large noun class systems, and agglutinative structures, where multiple morphemes are combined within a single word. In addition, frequent code-switching and orthographic variation further complicate text processing. These properties make token-level augmentation methods less accurate, since word boundaries and surface forms often fail to capture the underlying linguistic structure. This motivates our shift toward embedding-level augmentation, which can better generalize across morphological and orthographic variation.

This work not only addresses the pressing issue of data scarcity in Bantu languages but also provides a scalable and adaptable framework for extending text classification capabilities to other low-resource language families. Through the introduction of MAGE, we set the stage for future research in low-resource language processing and establish a pathway to improve the inclusivity and generalizability of NLP technologies.

2 Related Work

2.1 Data Augmentation Techniques

In recent years, data augmentation techniques have gained significant attention, especially for low-resource languages, due to the scarcity of properly annotated datasets and general lack of resources. Prior studies on data augmentation span several dimensions, and we group them here into three categories: lexical-level, contextual-level, and embedding-level augmentations.

Lexical-level augmentations One of the earliest and most widely used techniques is back-translation. Sennrich et al. [21] leverage monolingual target language data for textual-based augmentation using back-translation to enhance model performance, though at the cost of requiring an additional pretrained Neural Machine Translation (NMT) model. Lample et al. [8] propose a related method that relies solely on monolingual corpora by mapping sentences from two languages into a shared latent space via a shared encoder–decoder architecture. While these approaches reduce dependency on parallel corpora, they remain limited by the availability of monolingual data. Another influential lexical-level approach is Easy Data Augmentation (EDA) by Wei and Zou [24], which applies synonym replacement, random insertion, random swap, and random deletion. Despite its simplicity and reliance only on a synonym dictionary such as WordNet [11], EDA significantly improves model performance even when training on small datasets.

Contextual-level augmentations To overcome the limitations of predefined dictionaries, Kobayashi [7] proposed contextual augmentation, which leverages bidirectional language models to generate substitute words based on surrounding context. This produces more semantically appropriate alternatives and outperforms lexical-level methods like EDA, especially in low-resource settings. Generative adversarial models have also been applied in this space: Yu et al. [25] introduced SeqGAN, which combines reinforcement learning with GANs to generate discrete sequences. SeqGAN has been shown to improve fluency and diversity in sequence generation tasks such as NLP and music generation. Related adversarial approaches include Jia and Liang [5], who designed adversarial examples for reading comprehension, demonstrating performance drops even in state-of-the-art models on SQuAD [19]. Beyond augmentation, Raffel et al. [17] explored transfer learning for low-resource languages via sentence-level alignment and multilingual embeddings, and Li et al. [9] investigated synthetic data generation using large language models in zero- and few-shot settings. These contextual-level techniques extend augmentation beyond surface-level manipulation and exploit broader semantic and generative modeling.

Embedding-level augmentations More recent work explores perturbations directly in the representation space. Chen et al. [3] proposed TMix, which interpolates hidden representations of text samples, combined with entropy minimization and consistency regularization, to improve generalization in resource-limited settings. Such mixup-style strategies avoid the brittleness of token-level methods and are better suited to morphologically complex languages. Building on this line of work, we investigate embedding-level augmentation tailored to low-resource Bantu languages, where surface token manipulations often fail due to agglutination and orthographic variation.

2.2 Data Augmentation for Low Resource Corpora Text Classification

For downstream NLP tasks in low-resource settings, various augmentation methods have been proposed. Rahamim et al. [18] introduced TAU-DR, which employs soft prompts while keeping the language model frozen, reconstructing hidden representations into synthetic sentences. This improves multi-class classification without requiring additional model training. Thangaraj et al. [23] investigated cross-lingual transfer in African languages, benchmarking forgetting metrics, though without applying augmentation. Karimi et al. [6] proposed AEDA, which introduces punctuation marks into sentences as a lightweight augmentation, preserving semantic consistency and outperforming EDA in low-resource settings. Litake et al. [10] developed IndiText Boost, a framework designed for underrepresented Indian languages, which combines EDA and back-translation to outperform more complex LLM-based methods on classification tasks. Prompt-based augmentation has also been explored: Sahu et al. [20] used pretrained LLMs like GPT-2 [16] to generate synthetic intent-classification data, though reliability issues with LLMs sometimes degrade data quality. Zhao et al. [26] introduced EPiDA, an augmentation framework combining conditional entropy minimization with relative entropy maximization, balancing diversity and quality. EPiDA consistently outperforms earlier techniques across text classification tasks, highlighting its applicability for low-resource settings.

2.3 LiDA - Language Independent Data Augmentation

Sujana and Kao [22] proposed LiDA, a language-independent augmentation method for text classification. Instead of generating new sentences, LiDA perturbs sentence-level embeddings trained with multilingual SBERT, resulting in consistent gains of 2-3% in LSTM-based classification. Its language independence stems from the multilingual dataset used to train the underlying embeddings, making it broadly applicable to low-resource contexts.

3 Methodology

Given the morphological richness and agglutinative nature of Bantu languages, token-level augmentation risks introducing noise rather than diversity. For instance, splitting or replacing tokens without accounting for noun class agreement can distort meaning. Similarly, code-switching and orthographic inconsistencies challenge augmentation methods that assume stable token inventories. To address these issues, we adopt an embedding-level approach that operates on distributed representations rather than surface tokens, allowing us to capture linguistic variation more robustly.

Hence, further refining the LiDA architecture, we propose a multi-head attention-based mechanism to quantitatively highlight and weight the individual embeddings to emphasize the important contributions of the LiDA architecture for the text-classification goal.

3.1 Dataset

The dataset referred to is the AfriSenti SemEval Shared Task - 12 dataset by Muhammad et al. [12] based on tweet sentiment analysis. As the study focuses on the Bantu language family, the datasets of the following 3 Bantu languages were chosen - Kinyarwanda, Xitsonga and Swahili having 7940 tweet-label pairs in the combined training set and 1482 tweet-label pairs in the combined test set. We observe a skewness in the data towards Kinyarwanda due to Kinyarwanda having the highest data points at 5155 tweets, with Swahili being the second highest at 3009 tweets and Xitsonga having the least data at 1258 tweets.

Field	Description
ID	Alpha-Numeric Serial Numbers
Tweet	Tweet Content
Label	Tweet Label

Table 1. Field descriptions of the dataset.

The dataset consists of 3 main fields, namely ID, Tweet, and Label. On preprocessing the data, the ID field was dropped, and the tweets accordingly preprocessed: lowercasing and removing punctuations, hyperlinks, and emojis. The labels in the label field as mentioned in table 1 were given as Negative, Neutral and Positive to indicate the tweet sentiment which was label-encoded as 0,1,2 respectively for further processing.

3.2 Architecture

The previous architectures and frameworks discussed have focused primarily on widely studied languages such as English, Indonesian, Chinese, French, and others. Although these languages benefit from extensive resources and established linguistic frameworks, our work diverges by addressing African Bantu languages, which are linguistically distinct and underrepresented in computational research. Bantu languages exhibit unique structural and morphological characteristics, requiring specialized approaches that go beyond the methodologies applied to more commonly studied languages. Hence, the embedding models and the architectural complexities in the components of previous frameworks do not conform to the requirements of the Bantu languages.

Taking LiDA as our base framework, we propose our modified architecture in 3.2 creating a stable architecture that caters to the demands of Bantu languages.

LiDA Architecture The LiDA architecture (*figure 1*) makes use of the multilingual SBERT (Sentence-BERT) model making the architecture *language independent* in essence. The embeddings so generated are passed through three functions - linear transformation, autoencoder model, denoising autoencoder model - before being concatenated with the original embeddings and henceforth classified using LSTM and BERT classifiers.

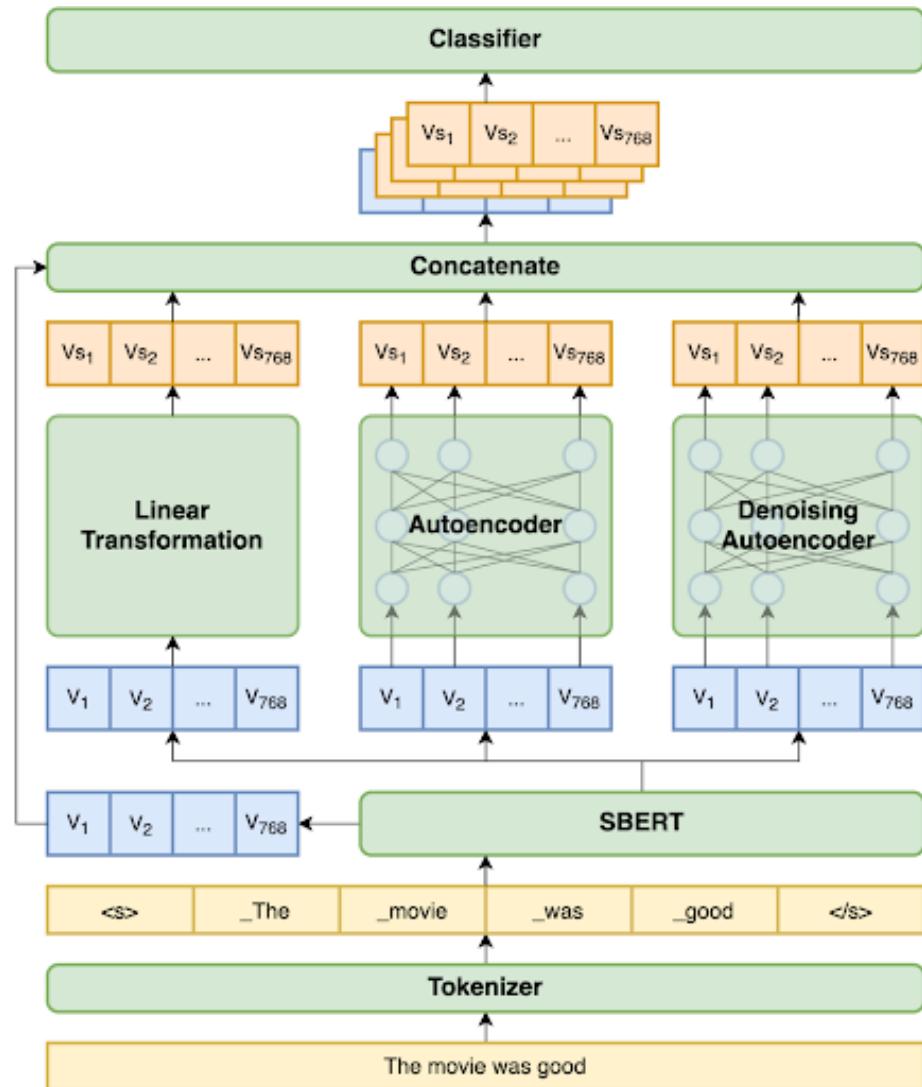


Fig. 1. LiDA Framework Reproduced from [22] [22].

Proposed Architecture Figure 2 shows our modification to the original LiDA architecture by changing the embedding model to AfriBERTa, replacing the denoising autoencoder with a variational autoencoder and addressing the concatenation of complex low-resource languages such as the Bantu language family by the introduction of weighted concatenation using multi-head attention. The tweets are passed through the AfriBERTa model, the choice of which is discussed in the sub-section 3.3. The model outputs a 768-dimensional representation of the text which is passed through the aforementioned transformation functions.

The Linear Transformation Layer introduces controlled variability into the input embeddings by applying a randomized shift, enhancing the robustness and generalizability of the representations. For each embedding, a random noise vector r is sampled uniformly within a range $[r_{\min}, r_{\max}]$ and added to the original embedding, resulting in a transformed embedding $e' = e+r$. This operation is performed independently for all embeddings in the training and testing datasets. The parameters r_{\min} and r_{\max} can be adjusted to control the magnitude of perturbation, ensuring that the embeddings retain their original semantic structure while introducing sufficient variability to aid learning.

Autoencoder is a key part of the augmentation process, designed to refine and diversify input embeddings by learning compressed representations while retaining essential features. This is accomplished through an encoder-decoder architecture that reduces the input embedding dimensions to a latent space and reconstructs them back to the original size. This introduces subtle variations while preserving essential semantic features, enhancing the diversity of augmented data. The model has been slightly enhanced from the original model used in Sujana and Kao ([22]) with Leaky ReLU activations, Batch Normalization - to ensure stable training and mitigate vanishing gradients - and Dropout layers to help with regularization, improve generalization and reduce overfitting for the Bantu language family. The encoder consists of sequential linear layers that progressively reduce the embedding size from the original 768 dimensions to 32 dimensions in the latent space with a learning rate of 0.001 for stabilized learning. The decoder mirrors the encoder structure, gradually increasing the dimensionality from the latent space back to the original embedding size of 768, with the final layer applying a sigmoid activation for bounded output. This structure ensures that the embeddings are refined through compression and reconstruction, creating more diverse and stable representations.

The Variational Autoencoder (VAE) proves to be integral to our data augmentation process, providing a more flexible and expressive approach to embedding refinement compared to the original denoising autoencoder. Unlike traditional autoencoders, VAEs model the input data as probabilistic distributions rather than deterministic mappings, allowing for more varied and stable synthetic embeddings. This probabilistic framework facilitates the generation of diverse augmented data points, which is crucial in low-resource language tasks like those involving Bantu languages. The VAE learns a distribution over the latent space, enabling the generation of new samples by sampling from the learned distribution, enhancing the diversity of the training data and helping the model

generalize better. The VAE architecture consists of an encoder and a decoder. The encoder first maps the input embeddings from their original input dimension of 768 to a latent space of dimension 256. The size of the intermediate layers, i.e., hidden dimension is set to 512, balancing capacity and complexity. Batch Normalization, ReLU activations, and Dropout (set to 0.2) are applied to improve stability, avoid overfitting, and ensure stable learning. The encoder then produces two outputs: the mean μ and log-variance $\log(\sigma^2)$ of the latent distribution. The reparameterization trick is applied to sample from this distribution, allowing gradients to propagate through the sampling process and enabling accurate training. Just as in the autoencoder, the decoder mirrors the encoder, first mapping the latent representation of latent dimension back to the hidden space, and then expanding to the original input dimension. The final output of the decoder is a probabilistic reconstruction of the original input. Also we are not proposing that VAE will always outperform the original DAE (De-noising Autoencoder) configuration mentioned in the LiDA architecture, but it may prove to be comparable or even superior in some scenarios, and both configurations will provide us with the desired end results when used along with attention mechanisms. Finally, the embeddings we get through these transformations are weighted and concatenated with the original embeddings using a multi-head attention mechanism section 3.4.

3.3 AfriBERTa as the Embedding Model

To evaluate the effectiveness of multilingual transformer embeddings for low-resource sentiment classification, we fine-tuned several candidate models—mBERT, XLM-R, AfriBERTa, and BantuBERTa—on three languages (Kinyarwanda, Swahili, and Tsonga) as well as on a combined dataset. Each model was trained and evaluated under identical experimental settings, and we report the classification metrics averaged over runs for accuracy, precision, recall, and F1-score.

From Table 2, AfriBERTa demonstrates superior overall performance across nearly all metrics and datasets. For Kinyarwanda, AfriBERTa surpasses the next-best model (BantuBERTa) by over 1% in accuracy, while for Swahili and Tsonga, it maintains the highest precision and balanced recall–F1 trade-off. On the combined multilingual dataset, AfriBERTa yields the best macro-average scores (0.5695 accuracy, 0.5803 F1), showing its adaptability to cross-lingual data and justifying its selection as the embedding model for all subsequent MAGE experiments.

3.4 MAGE

The multi-head attention component is designed to enhance the model's ability to focus on important embeddings by assigning distinct weights to different embeddings using multiple attention heads. This method allows for the dynamic selection of which embeddings have more influence on the final classification decision, effectively highlighting the critical features. The num_heads was set to 4 ensuring that multiple perspectives of the embeddings are captured simultaneously. The use of attention mechanism stemmed from the observation that

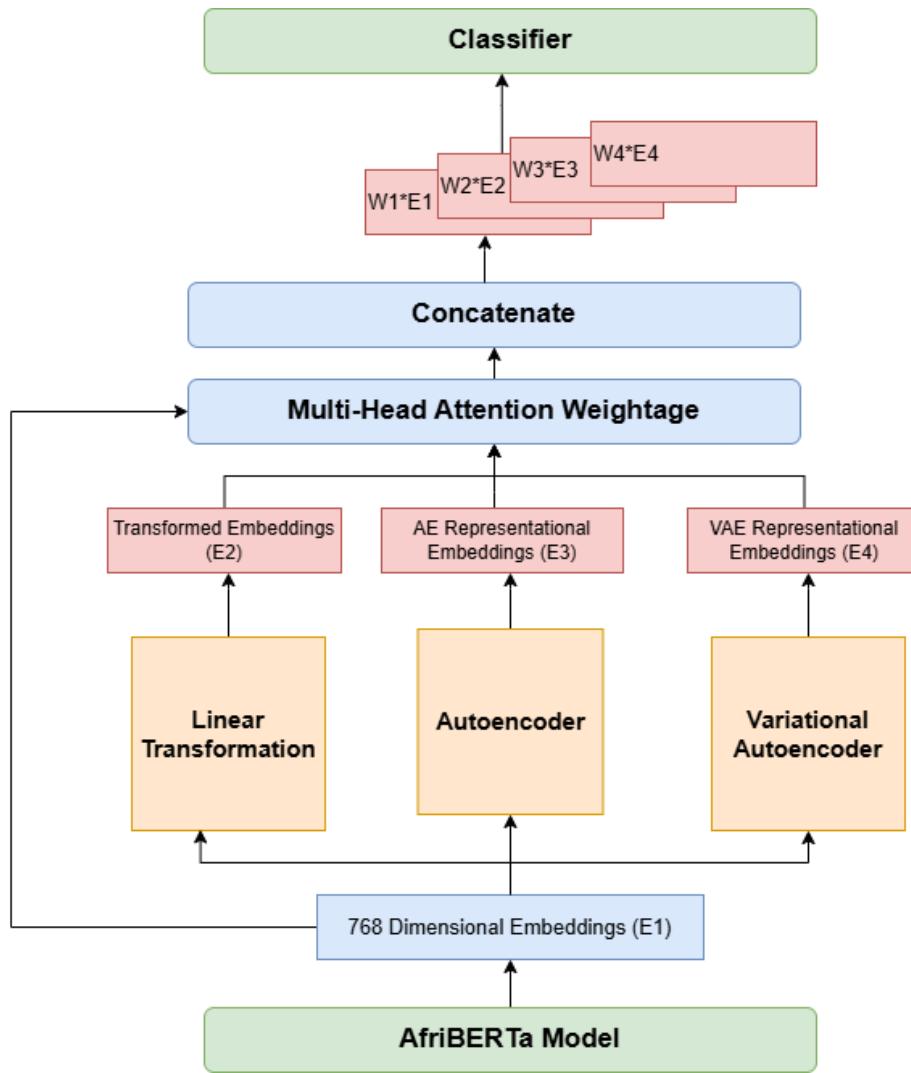


Fig. 2. Modified LiDA - MAGE Framework

Table 2. Language-wise classification performance of candidate embedding models. Metrics are averaged across experimental runs. AfriBERTa achieves consistently strong results across languages, outperforming others on the combined dataset.

Language	Model	Accuracy	Precision	Recall	F1
4*Kinyarwanda	mBERT	0.5296	0.5344	0.5196	0.5229
	XLM-R	0.3808	0.1269	0.3333	0.1838
	AfriBERTa	0.6408	0.6430	0.6391	0.6407
	BantuBERTa	0.6306	0.6346	0.6285	0.6309
4*Swahili	mBERT	0.5739	0.5193	0.4070	0.4052
	XLM-R	0.5960	0.4205	0.3405	0.2666
	AfriBERTa	0.6004	0.5427	0.4806	0.4966
	BantuBERTa	0.5960	0.5122	0.4869	0.4967
4*Tsonga	mBERT	0.5320	0.3569	0.4085	0.3692
	XLM-R	0.4729	0.1576	0.3333	0.2140
	AfriBERTa	0.5418	0.4892	0.4123	0.3772
	BantuBERTa	0.5172	0.4100	0.4099	0.3844
4*Combined Dataset	mBERT	0.4832	0.4640	0.4552	0.4250
	XLM-R	0.5207	0.5282	0.5497	0.5228
	AfriBERTa	0.5695	0.5785	0.5600	0.5803
	BantuBERTa	0.5266	0.5209	0.5113	0.5130

manually weighting embeddings improved classification results were observed, and attention provides a learnable way to optimize this process. In this approach, the embeddings are attended to using multi-head attention, where each head independently processes the embeddings and captures different aspects of the feature space. The context vectors, which are trainable, guide the attention mechanism to focus on the most relevant embeddings. The outputs from each attention head are then concatenated, and an aggregation is performed by summing the resulting vectors, capturing the most important features across different heads.

3.5 Classification

We employed two architectures for classification, namely, LSTM and logistic regression and the results were evaluated using accuracy, precision, recall and F1 score. The LSTM classifier was used with an input dimension of 768, hidden dimension of 128, and a single layer. The model was trained with a learning rate of 0.001 using CrossEntropyLoss. An early stopping mechanism with a patience of 3 epochs was employed, along with the StepLR scheduler to adjust the learning rate.

The second classifier used was logistic regression, which serves as a lightweight yet accurate baseline for classification. The model was trained with a maximum of 1000 iterations using the LBFGS(Limited-memory Broyden-Fletcher-Goldfarb-Shanno) solver for optimization. Since logistic regression is a simple

linear model, it provides a useful comparison against the LSTM’s sequential feature extraction capabilities. By analyzing both models, we aim to assess the impact of complex sequential modeling versus traditional linear classification on our dataset.

4 Results

In this section, we present the classification performance of various embedding configurations and attention mechanisms, evaluated using standard metrics such as accuracy, precision, recall, and F1 score. We further extended our experiments through multiple independent runs with shuffled datasets to assess the robustness and consistency of our proposed approach.

4.1 Effect of DAE and VAE

We first compare the performance of the original embeddings using the proposed VAE vs DAE configurations using LSTM and Logistic Regression without weighted attention concatenation mechanism.

Metric	Original	With DAE	With VAE
Accuracy	0.5680	0.5739	0.5769
Precision	0.5589	0.5659	0.5681
Recall	0.5554	0.5697	0.5562
F1 Score	0.5566	0.5672	0.5589

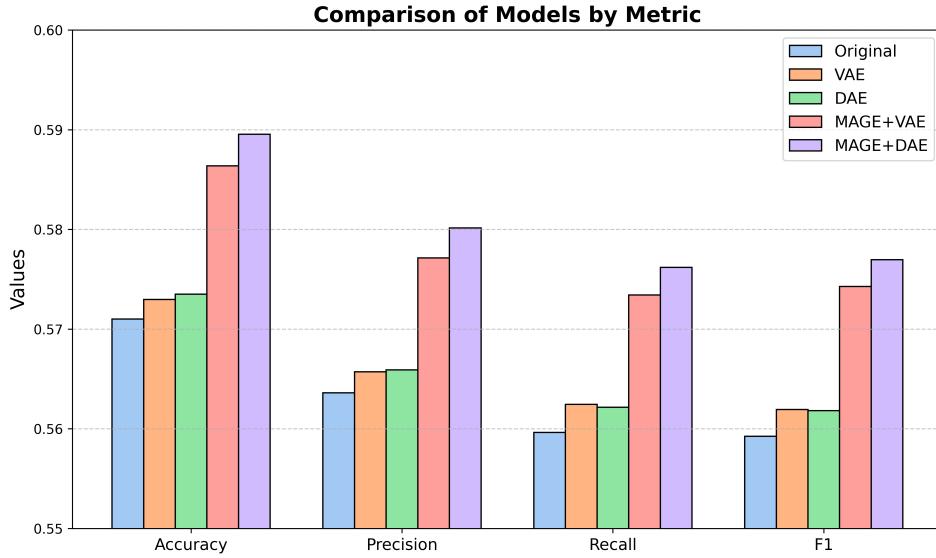
Table 3. Comparison between DAE and VAE configurations (LSTM).

Observing table 3, we see a consistent improvement in all metrics over the original results for both DAE and VAE configurations using the LSTM classifier.

Metric	Original	With DAE	With VAE
Accuracy	0.5710	0.5729	0.5735
Precision	0.5636	0.5672	0.5659
Recall	0.5596	0.5624	0.5621
F1 Score	0.5592	0.5619	0.5618

Table 4. Comparison between DAE and VAE configurations (Logistic Regression).

Similarly, for Logistic Regression in table 4, we observe for the VAE configuration a 0.34%, 0.37%, 0.50%, and 0.48% improvement in accuracy, precision, recall and F1 score respectively over the original classification results. The

**Fig. 3.** Comparative Results

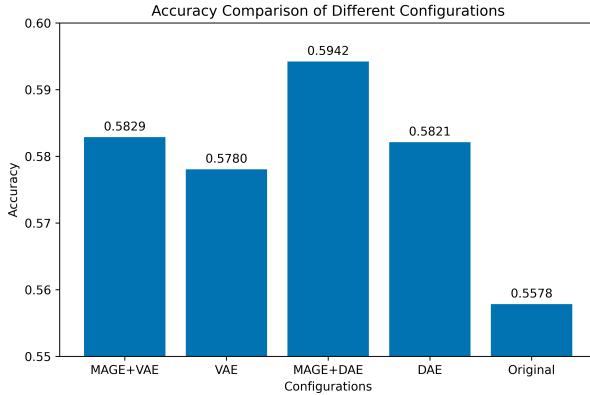
DAE configuration shows comparable improvements of 0.43%, 0.40%, 0.45%, and 0.46% for the same metrics. Both LSTM and Logistic Regression provide similar relative trends, suggesting that either autoencoder configuration can be effectively leveraged for feature enhancement.

4.2 Integrating Multi-Head Attention

To further test the robustness of our proposed attention-guided framework, we extended the experiments by introducing a Multi-Head attention mechanism. Each configuration was benchmarked by shuffling the dataset four times and running five independent training iterations per shuffle for both the LSTM and Logistic Regression classifiers. This approach allows us to evaluate performance stability and generalization beyond a single train-test split.

Figure 4 visualizes the average accuracy achieved across runs when comparing models with and without Multi-Head attention using both DAE and VAE configurations. The substantial performance gains are evident: **MAGE+DAE** achieves the highest accuracy improvement of 3.64% over the baseline, closely followed by **MAGE+VAE** at 2.51%. The attention mechanism further enhances both configurations, yielding an additional 0.4% and 1.21% increase over their respective non-attention counterparts.

A similar multi-run evaluation was conducted for the Logistic Regression classifier, where the averaged-out results (visualized in figure 3) exhibit consistent improvements across all metrics. To quantify these trends, we computed the overall mean and standard deviation (mean \pm std) across runs for both LSTM

**Fig. 4.** Comparative Results**Table 5.** Aggregated performance (mean \pm std) for Logistic Regression models averaged across 5 dataset shuffles and 5 independent runs.

Model	Accuracy	Precision	Recall	F1 Score
MAGE+VAE	0.5829 \pm 0.0078	0.5773 \pm 0.0065	0.5745 \pm 0.0069	0.5752 \pm 0.0070
VAE	0.5780 \pm 0.0060	0.5628 \pm 0.0054	0.5624 \pm 0.0057	0.5618 \pm 0.0058
MAGE+DAE	0.5942 \pm 0.0102	0.5839 \pm 0.0081	0.5803 \pm 0.0073	0.5805 \pm 0.0076
DAE	0.5821 \pm 0.0040	0.5631 \pm 0.0046	0.5612 \pm 0.0041	0.5613 \pm 0.0042
Original	0.5578 \pm 0.0132	0.5629 \pm 0.0109	0.5602 \pm 0.0107	0.5604 \pm 0.0109

and Logistic Regression models, aggregated across accuracy, precision, recall, and F1.

Table 6. Aggregated performance (mean \pm std) for LSTM models averaged across 4 dataset shuffles and 5 independent runs.

Model	Accuracy	Precision	Recall	F1 Score
MAGE+VAE	0.5864 \pm 0.0069	0.5802 \pm 0.0062	0.5791 \pm 0.0065	0.5805 \pm 0.0066
VAE	0.5795 \pm 0.0055	0.5734 \pm 0.0052	0.5706 \pm 0.0056	0.5720 \pm 0.0053
MAGE+DAE	0.5973 \pm 0.0088	0.5861 \pm 0.0074	0.5835 \pm 0.0071	0.5852 \pm 0.0072
DAE	0.5842 \pm 0.0048	0.5778 \pm 0.0046	0.5751 \pm 0.0047	0.5762 \pm 0.0048
Original	0.5591 \pm 0.0125	0.5613 \pm 0.0111	0.5585 \pm 0.0108	0.5588 \pm 0.0110

From these results, we observe that MAGE consistently improves performance across all four evaluation metrics for both classifiers. The performance gains are particularly pronounced in F1 score and precision, highlighting MAGE’s ability to better capture sentiment nuances. **MAGE+DAE** once again achieves

the best balance between accuracy and consistency, outperforming the baseline and single-autoencoder variants with low standard deviations across runs.

Overall, these experiments demonstrate that integrating Multi-Head Attention within the MAGE framework consistently improves classification accuracy, precision, recall, and F1 score across both LSTM and Logistic Regression models, validating its robustness and reproducibility across independent runs and shuffled datasets.

5 Conclusion

We thus present an innovative approach to embedding refinement and classification for the Bantu language family by integrating embedding-level transformations and high-performing attention mechanisms. Through systematic experimentation, we demonstrated that denoising and variational autoencoders enhance the quality of embeddings by refining their structure while preserving semantic integrity. A key contribution of this study was the introduction of a multi-head attention mechanism. The attention mechanism dynamically assigns different weights to various embeddings, enabling the model to focus on the most relevant features. This approach allowed us to capture crucial aspects of the embedding space more effectively. The results from our experiments clearly indicate that multi-head attention significantly boosted classification performance across various metrics. This validates the hypothesis that refining embeddings and emphasizing important features through attention enhances classification performance. The framework proposed in this study provides a scalable and stable solution, particularly for low-resource languages where linguistic diversity and data scarcity pose unique challenges.

6 Future Works

While this study demonstrates significant improvements in embedding refinement and classification through a novel architecture and attention mechanisms, it is limited to three Bantu languages. Testing on a broader range of Bantu and other low-resource languages is needed to assess the generalizability of the approach. Additionally, expanding the variety of pre-trained embeddings and exploring novel attention mechanisms, such as hierarchical or adaptive attention, could further enhance performance and applicability.

7 Limitations

While our proposed approach demonstrates notable improvements in embedding refinement and classification, it has several limitations. First, our experiments were conducted on a limited set of three Bantu languages, which restricts the generalizability of our findings to other Bantu and low-resource languages. Second, the dataset exhibits an imbalance, with Kinyarwanda comprising the majority of data points. This skewness may introduce biases in model learning and

affect the performance across languages. Third, the dataset size may be insufficient for training complex components like the Denoising Autoencoder, Variational Autoencoder, and the standard Autoencoder as they require a large and diverse dataset to learn meaningful latent representations effectively. The limited training data could lead to suboptimal embeddings, affecting downstream classification performance. Fourth, while embedding-level transformations using denoising and variational autoencoders refine embedding structures, their impact on preserving linguistic nuances requires further investigation. Finally, the computational complexity of our approach, particularly the attention mechanism, may pose challenges for real-time applications in resource-constrained environments. Addressing these limitations in future research will be essential for broader adoption and scalability.

References

1. Adelani, D.I.: Natural Language Processing for African Languages. Dissertation, Universität des Saarlandes (2022). <https://doi.org/10.22028/D291-40305>
2. Amol, C.J., Chimoto, E.A., Gesicho, R.D., Gitau, A.M., Etori, N.A., Kinyanjui, C., Ndung'u, S., Moruye, L., Ooko, S.O., Kitonga, K., Muhia, B., Gitau, C., Ndolo, A., Wanzare, L.D.A., Kahira, A.N., Tombe, R.: State of nlp in kenya: A survey (2024), <https://arxiv.org/abs/2410.09948>
3. Chen, J., Yang, Z., Yang, D.: Mixtext: Linguistically-informed interpolation of hidden space for semi-supervised text classification (2020), <https://arxiv.org/abs/2004.12239>
4. Jahan, M.S., Beddiar, D.R., Oussalah, M., Mohamed, M.: Data expansion using wordnet-based semantic expansion and word disambiguation for cyberbullying detection. In: Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022). pp. 1761–1770. European Language Resources Association (ELRA), Marseille, France (June 20–25 2022), <https://aclanthology.org/2022.lrec-1.187.pdf>
5. Jia, R., Liang, P.: Adversarial examples for evaluating reading comprehension systems (2017), <https://arxiv.org/abs/1707.07328>
6. Karimi, A., Rossi, L., Prati, A.: AEDA: An easier data augmentation technique for text classification. In: Moens, M.F., Huang, X., Specia, L., Yih, S.W.t. (eds.) Findings of the Association for Computational Linguistics: EMNLP 2021. pp. 2748–2754. Association for Computational Linguistics, Punta Cana, Dominican Republic (Nov 2021). <https://doi.org/10.18653/v1/2021.findings-emnlp.234>, <https://aclanthology.org/2021.findings-emnlp.234/>
7. Kobayashi, S.: Contextual augmentation: Data augmentation by words with paradigmatic relations (2018), <https://arxiv.org/abs/1805.06201>
8. Lample, G., Conneau, A., Denoyer, L., Ranzato, M.: Unsupervised machine translation using monolingual corpora only (2018), <https://arxiv.org/abs/1711.00043>
9. Li, Z., Zhu, H., Lu, Z., Yin, M.: Synthetic data generation with large language models for text classification: Potential and limitations (2023), <https://arxiv.org/abs/2310.07849>
10. Litake, O., Yagnik, N., Labhsetwar, S.R.: Inditext boost: Text augmentation for low resource india languages. ArXiv **abs/2401.13085** (2024), <https://api.semanticscholar.org/CorpusID:267200269>

11. Miller, G.A.: WordNet: A lexical database for English. In: Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994 (1994), <https://aclanthology.org/H94-1111/>
12. Muhammad, S.H., Abdulkumin, I., Yimam, S.M., Adelani, D.I., Ahmad, I.S., Ousidhoum, N., Ayele, A.A., Mohammad, S.M., Beloucif, M., Ruder, S.: SemEval-2023 Task 12: Sentiment Analysis for African Languages (AfriSenti-SemEval). In: Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023). Association for Computational Linguistics (2023)
13. Nie, E., Liang, S., Schmid, H., Schütze, H.: Cross-lingual retrieval augmented prompt for low-resource languages. In: Rogers, A., Boyd-Graber, J., Okazaki, N. (eds.) Findings of the Association for Computational Linguistics: ACL 2023. pp. 8320–8340. Association for Computational Linguistics, Toronto, Canada (Jul 2023). <https://doi.org/10.18653/v1/2023.findings-acl.528>, <https://aclanthology.org/2023.findings-acl.528/>
14. Ogueji, K., Zhu, Y., Lin, J.: Small data? no problem! exploring the viability of pretrained multilingual language models for low-resourced languages. In: Ataman, D., Birch, A., Conneau, A., Firat, O., Ruder, S., Sahin, G.G. (eds.) Proceedings of the 1st Workshop on Multilingual Representation Learning. pp. 116–126. Association for Computational Linguistics, Punta Cana, Dominican Republic (Nov 2021). <https://doi.org/10.18653/v1/2021.mrl-1.11>, <https://aclanthology.org/2021.mrl-1.11/>
15. Perçin, S., Galassi, A., Lagioia, F., Ruggeri, F., Santin, P., Sartor, G., Torroni, P.: Combining WordNet and word embeddings in data augmentation for legal texts. In: Aletras, N., Chalkidis, I., Barrett, L., Goantă, C., Preotiu-Pietro, D. (eds.) Proceedings of the Natural Legal Language Processing Workshop 2022. pp. 47–52. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Hybrid) (Dec 2022). <https://doi.org/10.18653/v1/2022.nlp-1.4>, <https://aclanthology.org/2022.nlp-1.4/>
16. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I.: Language models are unsupervised multitask learners (2019)
17. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer (2023), <https://arxiv.org/abs/1910.10683>
18. Rahamim, A., Uziel, G., Goldbraich, E., Anaby Tavor, A.: Text augmentation using dataset reconstruction for low-resource classification. In: Rogers, A., Boyd-Graber, J., Okazaki, N. (eds.) Findings of the Association for Computational Linguistics: ACL 2023. pp. 7389–7402. Association for Computational Linguistics, Toronto, Canada (Jul 2023). <https://doi.org/10.18653/v1/2023.findings-acl.466>, <https://aclanthology.org/2023.findings-acl.466/>
19. Rajpurkar, P., Zhang, J., Lopyrev, K., Liang, P.: SQuAD: 100,000+ questions for machine comprehension of text. In: Su, J., Duh, K., Carreras, X. (eds.) Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. pp. 2383–2392. Association for Computational Linguistics, Austin, Texas (Nov 2016). <https://doi.org/10.18653/v1/D16-1264>, <https://aclanthology.org/D16-1264>
20. Sahu, G., Rodriguez, P., Laradji, I., Atighechian, P., Vazquez, D., Bahdanau, D.: Data augmentation for intent classification with off-the-shelf large language models. In: Liu, B., Papangelis, A., Ultes, S., Rastogi, A., Chen, Y.N., Spithourakis, G., Nouri, E., Shi, W. (eds.) Proceedings of the 4th Workshop on NLP for Conversational AI. pp. 47–57. Association for Computational Linguistics (2023)

- tics, Dublin, Ireland (May 2022). <https://doi.org/10.18653/v1/2022.nlp4convai-1.5>, <https://aclanthology.org/2022.nlp4convai-1.5/>
21. Sennrich, R., Haddow, B., Birch, A.: Improving neural machine translation models with monolingual data (2016), <https://arxiv.org/abs/1511.06709>
 22. Sujana, Y., Kao, H.Y.: Lida: Language-independent data augmentation for text classification. *IEEE Access* **PP**, 1–1 (01 2023). <https://doi.org/10.1109/ACCESS.2023.3234019>
 23. Thangaraj, H., Chenat, A., Walia, J.S., Marivate, V.: Cross-lingual transfer of multilingual models on low resource african languages (2024), <https://arxiv.org/abs/2409.10965>
 24. Wei, J., Zou, K.: Eda: Easy data augmentation techniques for boosting performance on text classification tasks (2019), <https://arxiv.org/abs/1901.11196>
 25. Yu, L., Zhang, W., Wang, J., Yu, Y.: Seqgan: Sequence generative adversarial nets with policy gradient (2017), <https://arxiv.org/abs/1609.05473>
 26. Zhao, M., Zhang, L., Xu, Y., Ding, J., Guan, J., Zhou, S.: EPiDA: An easy plug-in data augmentation framework for high performance text classification. In: Carpuat, M., de Marneffe, M.C., Meza Ruiz, I.V. (eds.) Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 4742–4752. Association for Computational Linguistics, Seattle, United States (Jul 2022). <https://doi.org/10.18653/v1/2022.naacl-main.349>, <https://aclanthology.org/2022.naacl-main.349/>
 27. Şahin, G.G.: To augment or not to augment? a comparative study on text augmentation techniques for low-resource nlp. *Computational Linguistics* **48**(1), 5–42 (04 2022). https://doi.org/10.1162/coli_a_00425, https://doi.org/10.1162/coli_a_00425

An End-to-End Deep Learning Model for Recommender Systems

Abebe Tegene^{1,3}, Vukosi Marivate², Mapundi Banda², Abiodun Modupe¹, Valisoa Rakotonarivo², and Mathibele Nchabeleng²

¹ Department of Computer Science, University of Pretoria, Pretoria, RSA,
abebe.tegene@up.ac.za

² Department of Mathematics and Applied Mathematics, University of Pretoria, Pretoria, RSA

³ CoE of Big Data Analytics and HPC, Addis Ababa Science and Technology University, Addis Ababa 16417, Ethiopia

Abstract. The increasing availability of choices on online platforms has resulted in a rise in user expectations for personalized content across social media, entertainment, and e-commerce websites. Recommender systems, using machine learning, analyze user preferences to provide intelligent recommendations, help users manage information overload. Collaborative filtering, particularly based on latent factor models like matrix factorization, has proven successful in personalized recommendations but encounters challenges such as data sparsity and issues with non-linear feature representation. The use of deep neural networks for personalized recommendations has garnered interest to the advancements in deep learning methodologies. To improve recommendation performance and user experience, research is focusing on applying deep learning concepts to address existing challenges. This paper proposes an end-to-end deep learning framework to address these challenges. The fundamental idea of this method involves transforming the dense feature vector produced by embedding methods into two end-to-end deep neural network designs. Subsequently, it independently learns a low-dimensional feature representation and non-linear abstraction of the sparse data. In addition, the method incorporates a deep learning architecture into the output layer of the networks to predict the required rating scores. In four real-world datasets, this proposed technique surpassed state-of-the-art models in terms of performance.

Keywords: deep learning, matrix factorization, collaborative filtering, Recommender systems

1 Introduction

The diversity of choices available to consumers has increased dramatically as a result of the increasing use of the Internet. On modern social media, entertainment, and e-commerce platforms, consumers often find a wide range of well-liked products, movies, music, and restaurants. This results in an increasing volume of

content available on a given platform; thus, users frequently experience information overload, making it challenging to select anything suitable from an extensive number of options. As a result, personalized advice is the core strategy to provide clients with a better user experience [1, 2].

Recommender systems (RSs) have become widely recognized as an essential tool to help consumers navigate a wide range of options while also keeping them interested and happy with personalized content [1, 3]. RSs are classified into three distinct groups, based on the input data they use: collaborative filtering (CF), content-based (CB), and hybrid [4]. CF is an effective implementation strategy that makes recommendations based on previous user-item interactions, using an explicit or implicit interaction data matrix; however, it suffers from different issues: data sparsity, with non-linear feature representation, and cold start problems [5, 6]

Recent advances in deep learning (DL) have produced state-of-the-art results in a variety of fields, including image recognition, natural language processing, computer vision, RSs, and many more [7, 8]. The DL method can accomplish automatic feature extraction, while traditional RSs techniques require human feature extraction [9]. Because there are more users and items than ever before, the RS issues have high-dimension data. Thus, DL can be used to automatically extract high-dimensional rich feature representations from these data [10, 11]. Even though DL has had some successful results in other domains. Those issues are still exist in RSs [7]. Therefore, it has become essential to find a more robust method of applying DL techniques in the field of RSs to address the aforementioned problems and improve their performance.

In this work, we explore the benefits of the DL framework and introduce a DL method to enhance the performance of RSs. We introduce an end-to-end DL method for these systems, referred to as EDRS. To enhance the effectiveness of recommendations, EDRS uses embedding and DNN (Deep Neural Network) structures to acquire the hidden latent features. The model combines deep learning techniques to extract complex abstractions and nonlinear feature representations from the data for predicting ratings. The core idea is to transform the dense user and item vectors generated by the embedding methods into two fully connected DNN architectures.

In general, in most existing RS designs that utilize DL, user and item input vectors are combined and processed to pass through a shared DNN to extract features. We believe that this approach lacks the resilience needed to effectively capture the necessary information from the data. Due to the intricate nature of user behaviors, we recommend analyzing these aspects separately to acquire the relevant insights. We contend that the current strategy falls to robustly learn the essential features from the data. Previous methods have not differentiated the architectures for these distinct components. In contrast, the proposed technique assigns the input vectors of the user and the item to its fully connected deep neural networks. Additionally, the method integrate a multi-layer perceptron (MLP) architecture into the proposed method to predict the desired rating scores. This model architecture effectively captures the intricate features

of user and item within their respective frameworks, suggesting that utilizing this proposed end-to-end architecture will significantly improve the performance of recommendations.

To summarize, the key contributions of the proposed work are as follows:

- We developed a DL architecture to factorize a user-item interaction matrix. This approach involves learning the non-linear and the abstract hidden features of users and items within user and item deep learning structures.
- We developed an end-to-end DL framework for recommender systems that effectively addresses the sparsity issues.
- We incorporated an MLP network into the output layer of the proposed architectures to predict rating scores. This integration demonstrates that the techniques significantly enhance recommendation quality, achieving state-of-the-art performance levels.

2 Related Work

This section reveals pertinent research studies about the current study of recommender systems.

2.1 Collaborative Filtering

The CF-based recommendation method works under the fundamental premise that users would select products that are similar to themselves if they had similar interests. Thus, identifying users who share the target user's interests and preferences is crucial for CF-based RSs [12, 13]. For instance, probabilistic matrix factorization (PMF) is one of the CF approaches the model user-item rating matrix as a product of two low-dimensional latent factor matrices using Probabilistic framework to predict missing rating. This method overcomes sparsity problems [14].

One of the efficient latent factor techniques in the CF method is the matrix factorization (MF) model. The method decomposes a high-dimensional data matrix into the product of two new low-rank matrices. This technique maps the latent feature components towards a common latent space. Next, user preferences for items in this space are predicted using the dot product within the latent feature vectors of the user and item [15, 16, 17].

$$\sum_{i,j} e_{ij}^2 = \min_{p_i, q_j} \sum_{i,j} (r_{ij} - \hat{r}_{ij})^2. \quad (1)$$

In most MF models, the latent factors p_i and q_j are typically calculated using stochastic gradient descent methods to optimize the loss given in (Eqn. 1) [17, 18, 19]. In this regard, SVD [17, 18], is one of the best MF techniques available to improve scalability and reliability problems. For example, biased SVD included biasing characteristics in the model to enhance rating prediction [18]. The SVD++ method produced good prediction performance by breaking down

the rating interaction matrix into low-rank matrices [19]. In addition, Regularized Matrix Factorization (RMF) overcomes overfitting and improves generalization in RS tasks by assigning regularization terms to the users and item latent factors [20].

All of these studies demonstrate the value of the MF technique as a foundational tool for RSs and highlight its numerous benefits for mixing different types of data. However, these approaches lack robustness, since they need too many iterations to find appropriate latent characteristic representations. To avoid this, Rendle, Steffen [21] recommended utilizing a large dimension. However, this also affects the quality of the recommendation in the event of a sparse data matrix. Therefore, the proposed approach is well suited to solve the problem.

2.2 Deep Learning Methods

Recent developments in DL-based recommendations have received great interest due to their significant ability to surpass the limitations of conventional methods and generate exceptionally high-quality recommendations [7, 22]. These capabilities enable DL-based RSs to improve user satisfaction

To generate an end-to-end model, the DeepFM [23] combines the factorization machine (FM) and MLP. This system's DNN captures the data's abstract representation. The FM employs two operations, addition and inner product to capture pairwise and linear interactions between features. Mongia et al. proposed a deep latent factor model(DLFCF) that uses multiple nonlinear layers to capture the complex user-item interactions. The method showed better performance over traditional MF methods [24]. Xue, Hong-Jian, et al also developed a deep matrix factorization (DMF) model that extends traditional MF models to learn nonlinear interaction of users and items latent factors using DNNs, and showed astonishing performances [15].

Alashkar, Taleb, et al. proposed a cosmetics recommendation model using MLP [25]. The method employs expert rules and labels instances using two identical MLPs. Using MLP and MF approaches, the NN architecture can be used within the inner product framework. For example, in NMF [26], a feed-forward neural network takes on the role of the classic inner product. In Neural Collaborative Filtering (NCF) framework, He, Xiangnan, et al. integrate the output of an MLP coupled with the latent elements of MF using a nonlinear transformation, before making the final recommendation, [11].

Abebe et al. proposed a deep learning-based matrix factorization method by extending traditional MF using DNNs. The model captures complex and high-order user and item interactions, and improved the accuracy of the recommendation much better than MF models [27, 28]. Zheng et al. proposed a convolution neural network (CNN) to learn the latent representation of users and items. By capturing local patterns, the method improves the accuracy of the rating prediction [29]. Adaptive deep RS (ADRAS) further adapts to both rating and ranking prediction tasks, showing improved performance due to its adaptive strategy [30].

As illustrated, the fundamental issues with the RS tasks are: a significant amount of data is shared between users and items; sparsity; and handling non-linear interaction problems. These results make feature learning a challenging task. Therefore, to address the drawbacks, this paper proposes an end-to-end deep learning method to discover the latent factors. It also integrates embedding approaches to address sparsity issues in the user-item interaction data matrix. In conclusion, enhancing the performance of recommender system models through DL methods is the goal of this paper.

3 Methodology

3.1 Problem Definition

Let $P = \{p_1, p_2, p_3, \dots, p_m\}$ and $Q = \{q_1, q_2, q_3, \dots, q_n\}$ represents the sets of m users and n items in the matrix, respectively. Let $Y \in \mathbf{R}^{m \times n}$ be the rating matrix, where $s_{ij} \in Y$ denote the rating given by user i to item j . The user-item interaction rating matrix I is then constructed, using (Eqn. 2).

$$r_{ij} = \begin{cases} s_{ij} & \text{if } s_{ij} \in Y \\ \text{unknown} & \text{otherwise} \end{cases} \quad (2)$$

3.2 General Architecture EDSRS

The ultimate goal of the proposed method is to investigate whether DL-based prediction is superior to that of simple dot products. To this end, an end-to-end DL architecture is proposed and depicted in Fig. 1.

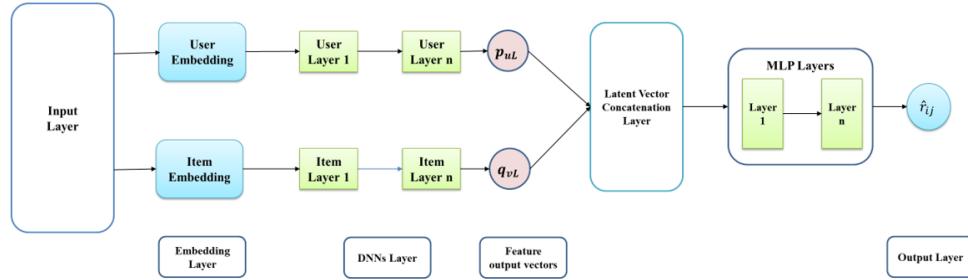


Fig. 1: The architecture of EDRS.

The method has two separate DNN architectures for user and item latent feature representation leaning. The approach integrates a dual DL and embedding techniques to extract abstract and non-linear characteristic representations

from the interaction information for the rating prediction task. That is, the latent factors from the two deep neural network architectures are merged and transferred into an MLP layer to jointly extract and predict the rating scores. That is, to solve the problems associated with latent factor models, such as sparsity, as well as the capacity to elicit deep abstract non-linear characteristics from data, the proposed method employs an end-to-end DL framework in the proposed strategy that incorporates embedding techniques to tackle those difficulties. The basic premise of the approach is mapping the dense latent vectors produced by embedding methods into two fully integrated user- and item-DNN structures. The rating value is then estimated using the DL method called MLP in the output layer of the network between the results of the two latent factors generated by the two structures. The details of the proposed architecture are discussed as follows:

Input layer: The proposed model's input layer takes explicit feature vectors that are derived from user interactions with items.

Embedding Layer: The EDRS model's embedding layer creates a dense, low-dimensional representation from high-dimensional input feature and sparse data. The basic feature vectors for user p and item q can be obtained in this embedding space in the manner described below:

Let for k embedding dimension, $P \in \mathbf{R}^{m \times k}$ be user embedding matrix and $Q \in \mathbf{R}^{n \times k}$ be the item embedding matrix that captures the latent features of users and items, respectively. The proposed method factorizes the rating matrix R into two low-rank matrices that satisfy $R \approx PQ^T$: using the loss function L in Eqn. (3) from known interaction ratings.

$$L = \min_{P, Q} \sum_{(i, j) \in R_{known}} (r_{ij} - p_i q_j^T)^2. \quad (3)$$

DNN Layers: This approach involves separately assigning the dense latent factor of user and the dense latent factor of item to two DL frameworks, allowing the model to capture nonlinear patterns and intrinsic abstractions within the dataset. In the user architecture of the proposed method, Eqn. (4) is used. In the equation $w_{u1}, w_{u2}, \dots, w_{uL}$ and $b_{u1}, b_{u2}, \dots, b_{uL}$ represent user weights and biases, where as σ in this context denotes an activation function, which in our case was *ReLU*.

$$\begin{aligned} h_{u1} &= \sigma(w_{u1}^T h_{u0} + b_{u1}), \\ h_{u2} &= \sigma(w_{u2}^T h_{u1} + b_{u2}), \\ &\vdots \\ h_{uL} &= \sigma(w_{uL}^T h_{uL-1} + b_{uL}), \end{aligned} \quad (4)$$

From each layer, we have an output $h_{u1}, h_{u2}, \dots, h_{uL}$. Similarly, we obtained $s_{u1}, s_{u2}, \dots, s_{uL}$ for the output layers of the items.

3.3 Problem Formulation for Prediction Layer

The proposed method utilizes a mapping $\phi : \mathbf{R}^k \times \mathbf{R}^k \rightarrow \mathbf{R}$ that combine two k -dimensional vectors $p \in \mathbf{R}^k$ and $q \in \mathbf{R}^k$ into a single score. For instance, if p denote the latent factor of a user u and q denotes the latent factor of an item v , then $\phi(p, q)$ represents the affinity of the user u to an item v .

In the proposed method, the h_{uL} and s_{uL} are the output features of a dual deep neural network architecture. This is used to predict the rating scores. Then, the dot product between the two factors can be defined using ϕ_{dot} :

$$\phi_{dot} = \langle h_{uL}, s_{uL} \rangle = h_{uL} s_{uL}^T \quad (5)$$

According to [31], MLPs are recognized as universal approximators that can estimate any continuous function in a compact set, provided that they possess an adequate number of hidden states. The following shows how a one layer MLP works:

A one layer MLP function $f : \mathbf{R}^s \rightarrow \mathbf{R}^t$ is defined as:

$$f_{W,b}(x) = \sigma(Wx + b) \quad (6)$$

where $W \in \mathbf{R}^{s \times t}$ and $b \in \mathbf{R}^t$ are parameters and σ is an activation function. MLP is a device that stacks many layers of function f as a composition for approximations. For example, a four layer MLP would have this structure,

$$f_{W_4,b_4}(f_{W_3,b_3})(f_{W_2,b_2}(f_{W_1,b_1}(x)) \quad (7)$$

Consequently, the proposed method merged the two latent factors as $\langle h_{uL}, s_{uL} \rangle$ and applied an MLP function to predict the rating scores as follows:

$$\phi_{MLP} = f_{W_n,b_n}(\dots, (f_{W_1,b_1}([h_{uL}, s_{uL}]), \dots) \quad (8)$$

Nowadays, it has become popular to replace the dot product with DNN architectures. Most frequently, several scholars used MLP for their proposed architecture [11, 26, 32]. The reasoning behind this is that since MLPs are universal function approximators, they rigorously outperform fixed similarity functions like the inner product. This indicates that the proposed approach supports the hypothesis of the researchers.

Once the latent features h_{uL} and s_{uL} for users and items are acquired from the two DNN architectures, either of the following two equations is used to predict the target rating score, \hat{r}_{ij} .

$$\hat{r}_{ij} = \phi_{MLP} = f_{W_n,b_n}(\dots, (f_{W_1,b_1}([h_{uL}, s_{uL}]), \dots) \quad (9)$$

$$\hat{r}_{ij} = \phi_{dot} = h_{uL} \odot s_{uL}, \quad (10)$$

where \odot represents the inner product.

A crucial element in optimizing a recommendation model is having a well-designed objective function. We employ the mean squared error loss function to

optimize model parameters, as it has shown excellent performance with explicit point-wise data types [15].

$$L = \min_{p,q} \sum_{i,j} (r_{ij} - \hat{r}_{ij})^2 + \lambda W^2 + \alpha b^2. \quad (11)$$

where λ and α are regularization parameters for the weights and biases, respectively.

4 Experiments

4.1 Experimental Datasets

This research uses openly accessible MovieLens⁴ datasets provided by the GroupLens research team. These datasets consist of movie ratings of different sizes and are widely used to evaluate CF algorithms [11]. For comparing models, we selected a dataset with variants of $100K$ and $1M$. In addition, the paper used an Amazon dataset. The Amazon dataset⁵ is the largest RS dataset that has been widely used in various related works [30] (see Table 1).

Table 1: MovieLens and Amazon datasets.

Datasets	No. of users	No. of items	No. of ratings	Sparsity
ML-100K	944	1683	100,000	93.7%
ML-1M	6040	3706	1,000,208	95.5%
Amazon Instant Videos	5100	1596	37,132	99.5%
Amazon Musical Instruments	1498	1022	11,206	99.3%

4.2 Evaluation protocol

To ensure the correctness of the recommendation result, the models' estimating performance is examined using MAE and RMSE [11].

4.3 Baseline Methods

The proposed model is evaluated with the following baselines [24]

- SVD++ is a variation on biased SVD. It enhances the biased SVD model by including implicit information. Implicit feedback, such as surfing and purchasing history, might suggest user preferences when explicit feedback is unavailable. It produced a better recommendation results than regular SVD or biased SVD [33].

⁴ <https://files.grouplens.org/datasets/movielens/>

⁵ <http://jmcauley.ucsd.edu/data/amazon>

- PMF is a MF method that incorporates probabilities. It uses MF and probabilistic models to generate predictions based on given data, It provides a more robust and accurate model for a variety of applications [14].
- NCF: The NCF paradigm improves the recommendation process by using neural networks. It captures intricate user preferences and item information, allowing it to provide personalized recommendations [11].
- DLFMF is a latent factor-based CF model. To find their latent representation vectors, the model applies deep factorization to both users and items [24].
- DMF is a Deep MF model that leverages MLP to transform the user and item representation [15].
- RMF: The model minimizes the distances within the matrices by using the Manhattan distance. The issue of data sparsity is effectively resolved [20].
- DCF is a MF technique for CF that is based on DL. To learn latent features of user and item in their respective structures, it created a dual DNN architecture [27].
- CoNN Deep: [29] This method used CNNs to learn latent factor representations of the user and the items for the rating prediction task.
- ADRS model [30] is an adaptive DL-based model that uses DL techniques to address rating and ranking prediction tasks in RRs.
- DELCR is a DL and embedding-based method for CF. It uses a dot product in the output layer of the network to predict the rating score [28].

4.4 Parameter Settings

Training data makes up 70% of the data, while test and validation data make up the remaining portion. A representative prediction result is obtained by selecting the average value from five distinct training runs, as indicated in Table 2. Tensorflow⁶ is used in simulating the model.

For model optimization, Adam [34] is used with a batch size of 512, a learning rate of 0.0001, and a regularization coefficient λ of $1e^{-6}$. For dual DNN, the technique uses two hidden layers. The number of factors in hidden layers is 80, 40 for the user and for the items, while the number of embedding dimensions is set to 10. In the final layer of the MLP, two hidden layers are used with 80 and 40 neurons, respectively, along with a *ReLU* activation function in the output layers.

4.5 Experimental Results

The performance result and the percentage improvement result of the proposed method, are summarized in Table 2 and Table 3, respectively. The bold result is achieved using the proposed method. From the result, it can be inferred that the method gives a significantly improved overall performance result for all baseline methods.

⁶ <https://www.tensorflow.org>

Table 2: Model performance result for ML-100K and ML-1M dataset.

Methods	ML-100K		ML-1M	
	MAE	RMSE	MAE	RMSE
SVD++	0.736	0.936	0.703	0.888
PMF	0.728	0.920	0.695	0.873
NCF	0.755	0.961	0.695	0.873
RMF	0.732	0.938	0.689	0.876
DMF	0.735	0.940	0.691	0.878
DLCFC	0.717	0.901	0.678	0.854
EDRS	0.626	0.831	0.605	0.801

Table 3: The Percentage improvement result in EDRS for both datasets.

Methods	ML-100K		ML-1M	
	MAE	RMSE	MAE	RMSE
SVD++	14.95%	11.22%	13.94%	9.80%
PMF	14.01%	9.67%	12.95%	8.25%
NCF	17.09%	13.53%	12.95%	8.25%
RMF	14.48%	11.41%	12.19%	8.56%
DMF	14.83%	11.60%	12.45%	8.77%
DLCFC	12.69%	7.77%	10.77%	6.21%

5 Discussion

5.1 Performance Comparison

The figures referenced, namely Fig. 2a, and 3a, demonstrate that the suggested model achieves convergence across all datasets. In summary, here are the key points of the experiments:

As shown in Table 3, the proposed method, EDRS, surpasses all other approaches in all datasets. Specifically, compared to the highly regarded DMF model, the EDRS method achieves a relative enhancement of 14.30% in the mean absolute error and 11.60% in RMSE for the variation of 100K dataset. Moreover, it demonstrates an improvement of 12.45% in MAE and 8.77% in RMSE for the 1M dataset. Furthermore, compared to other DL-based models, the advances achieved by the proposed model are substantial. Similarly, the proposed method outperforms the NCF model, which utilizes an MLP in the output layer by a significant margin. This improved performance is due: (i) the utilization of embedded ratings, which effectively addresses the issue of sparse dataset; (ii) separate mapping of the dense embedded latent features to the proposed architecture aids in effectively uncovering the complex and abstract feature representation of users and items; and (iii) incorporation of an MLP layer in the output layer leads to a strong recommendation performance. The difference between EDRS and MF-based methods is once again quite significant. Compared to the 100K dataset, the EDRS method showed a 14.01% improvement in MAE and an 11.41% improvement in RMSE over the RMF method. Similarly, for

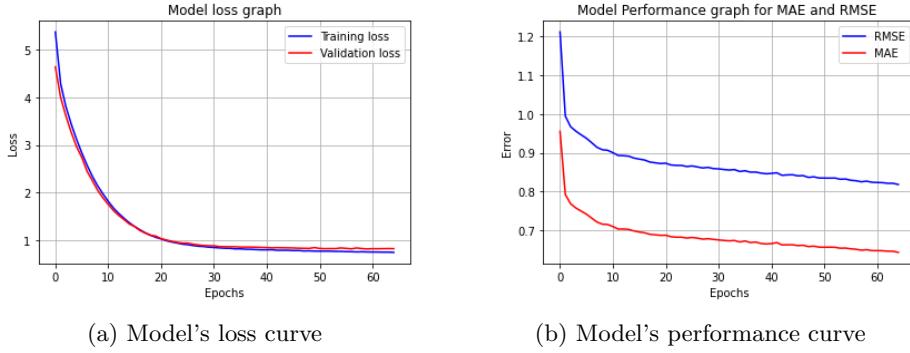


Fig. 2: The MAE and RMSE results of EDRS for MovieLens 100K dataset.

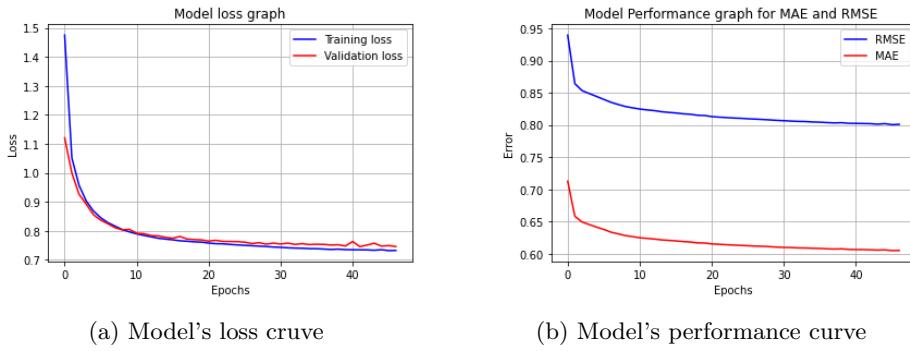


Fig. 3: The MAE and RMSE results of EDRS for MovieLens 1M dataset.

the ML-1M dataset, there was a an improvement of 12.19% in MAE and 8.56% in RMSE. The proposed technique outperforms the other MF methods, such as SVD++, by a considerable margin when compared to robust matrix factorization (RMF) methods. This demonstrates that utilizing an end-to-end deep learning architecture enables the method to effectively capture the underlying feature representation in a sparse dataset, which was a significant limitation of the traditional MF-based models.

The EDRS method, also evaluated based on the latent factor models DLFCF and DCF, shows relatively minimal improvement compared to the other methods. This is not surprising given that the model used a latent factor strategy. However, the improvement seen in the proposed model is attributed to its effective architecture to extract latent features in a robust manner. This shows that the proposed method, EDRS addresses the issue of data sparsity. Overall, EDRS performs well in terms of improving recommendation performance in all datasets. The improvement in performance underscores the effectiveness of the architecture in the method.

Table 4 shows the performance of the EDRS method with respect to the Amazon dataset. As shown, the EDRS model outperforms all baseline approaches, including the Deep-CoNN method. The improvements achieved by EDRS over baselines are significant. This performance improvement shows the effectiveness of the EDRS model compared to existing DL-based recommendation methods.

Table 4: The performance result of EDRS for the Musical Instrument and Instant Video Amazon dataset.

Methods	Musical Instruments		Instant Videos	
	MAE	RMSE	MAE	RMSE
PMF	0.982	1.004	0.907	1.101
Deep-CON	0.786	0.806	0.691	0.878
ADRS	0.525	0.766	0.505	0.757
EDRS	0.515	0.751	0.501	0.696

By evaluating the proposed models on these diverse datasets, spanning the e-commerce, review, and movie domains, we were able to gain a more comprehensive understanding of the models' performance characteristics and their ability to generalize across different contexts. The results of the expanded evaluation showed that the models maintained strong performance in the various datasets, demonstrating their robustness and adaptability to different data characteristics and use cases. This provides greater confidence in the effectiveness of the approaches and their potential for practical deployment in real-world scenarios. Overall, the expanded evaluation has strengthened the findings of the chapter and provided valuable insights into the generalizability and practical applicability of the proposed recommender system models.

5.2 Impact of Dual Deep Learning Layer

The proposed EDRS model is a novel deep learning architecture that uses an end-to-end dual deep learning architecture, which is a significant technical contribution to the field of recommender systems. To expand on this innovation, we provide a comprehensive overview of model architecture and the rationale behind the dual-layer design. When comparing EDRS with one of the representative deep learning-based baseline NCF model both utilize a Multi-Layer Perceptron (MLP) in the output layer, but as observed in Fig. 4 EDRS leverages a Dual Deep Learning Layer, which enables it to significantly outperform the NCF baseline.

An important distinction between EDRS and DELCR is the output layer design. DELCR uses an inner product operation, which likely enables more direct modeling of feature inter-actions, compared to the generic MLP used in EDRS and NCF. The consistently superior performance of both the EDRS and DELCR models, achieved through the innovative dual Deep Learning Layer architecture,

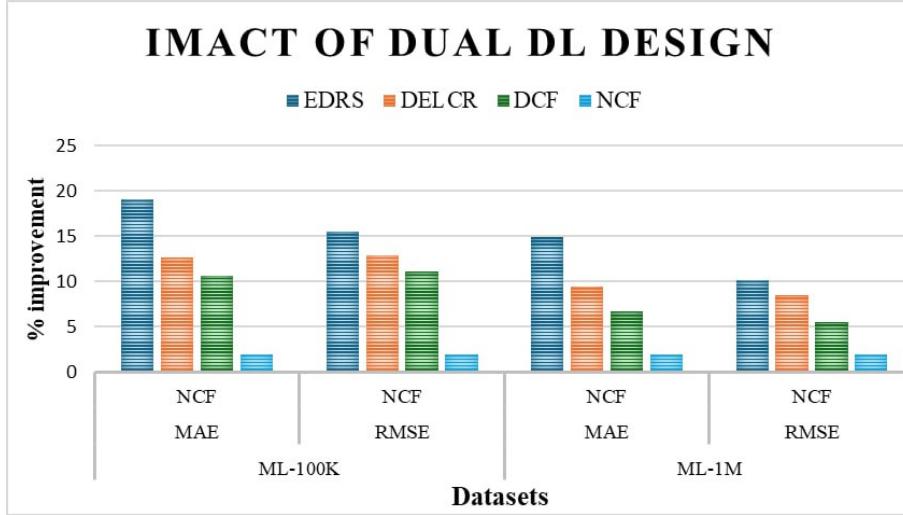


Fig. 4: The architecture advancement impact of the proposed model EDRS.

Table 5: The Performance result of EDRS based on MLP over dot product for ML-100K and ML-1M datasets.

Models	ML-100K		ML-1M	
	MAE (%)	RMSE (%)	MAE (%)	RMSE (%)
DCF	0.698 (+10.32)	0.887 (+6.31)	0.662 (+8.61)	0.842 (+4.87)
DELCR	0.674 (+7.12)	0.856 (+2.92)	0.643 (+5.91)	0.818 (+2.08)

highlights the value of exploring complex deep learning designs to unlock better performing models, which is a significant contribution to the field of deep learning-based recommender systems.

From this we can infer that the innovation of the dual deep learning layer architecture is a significant contribution of the proposed model to the field of deep learning-based recommender systems. This architectural design, which adds an extra deep learning layer compared to baseline models, is a core technical advancement that enables the model to learn more sophisticated representations and achieve substantial performance improvements over existing approaches. The emphasis on the dual deep learning layer as the primary innovation and technical contribution of the proposed model highlights how advancements in deep learning architecture design can drive meaningful progress in developing more effective recommender systems.

5.3 Dot Product vs MLP

The primary goal of this section is to determine whether MLP-learned similarity is a better option than a simple dot product. As can be seen in Table 5, an

MLP based model substantially outperforms the dot product based baselines on all datasets. For instance, the relative improvement achieved by EDRS using MAE over DCF and DELCR for the ML-100K dataset is 10.32% and 7.12%, respectively. This result demonstrated that EDRS achieved the best performance in improving recommendation performance. The experimental results validate the claim in [11] that a dot product model's embeddings through an MLP can improve the model. The findings of the variants of NeuMF [11] further support the proposition that a learned similarity via an MLP is better than a dot product. In conclusion, the findings confirm that a learned resemblance with an MLP is better than a dot product.

5.4 Ablation Study

To explore the impact of different elements of the EDRS model, this section presented an ablation study by testing the model in various configurations. Figure 5 shows the ablation results.

EDRS-Default represents the baseline configuration of the method presented in this paper. This variation is specifically designed to assess the impact of incorporating DL in the network output layer, achieved by replacing the inner product with a multilayer perceptron.

DELCR : This variation is specifically designed to assess the impact of incorporating dual DL in the network and the use of the inner product in the output layer of the architecture.

DELCR-WOE (without Embedding): This variant was created to evaluate the influence of the embedding techniques. In this configuration, the embedding network is excluded, and the sparse user-item interaction matrix is filled directly with zero values. EDRS-WOE (without Embedding): This version is specifically

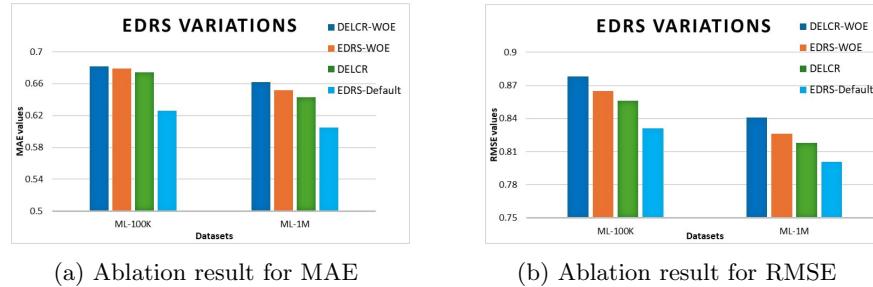


Fig. 5: The Ablation result of the proposed mdel using MovieLens dataset.

tailored to assess the impact of embedding techniques in EDRS. In this setup, the embedding network is omitted, and the sparse user-item interaction matrix is directly populated with zero values.

As depicted in Fig. 5, the different configurations of the EDRS method demonstrate varying performance levels in diverse datasets. The EDRS-default,

which represents the standard configuration of the proposed approach, generally outperforms other setups such as DELCR-WOE, EDRS-WOE and DELCR. This emphasizes the substantial influence of end-to-end deep learning strategies on performance improvement.

The EDRS-default shows superior performance, emphasizing the crucial role of embedding techniques and deep learning methods in boosting network performance. In contrast, the EDRS-WOE setup, which omits the embedding network and directly incorporates zero values into the sparse user-item interaction matrix, still achieves good results. This demonstrates the effectiveness of the multi-layer perceptron in improving model performance and underscores the beneficial influence of embedding strategies on the accuracy of recommendation systems.

Furthermore, the DELCR version demonstrates relatively strong performance because it utilizes both dual deep learning and embedding techniques within the network. This highlights the importance of these approaches for enhancing the performance of recommendation systems.

Moreover, the DELCR-WOE version shows low performance due to its sparse user-item interaction matrix resulting from the direct insertion of zero values into the sparse user-item interaction matrix. This indicates that neglecting embedding approaches reduces model accuracy, emphasizing the importance of these strategies for improved recommendation system performance.

6 Conclusion and Future Works

In this study, we investigated an end-to-end deep learning approach for recommender systems. The method involves initially embedding the input vectors of users and items to create dense low-dimensional representations that are then individually fed into two DNN architectures to unveil abstract and nonlinear data representations. Subsequently, the outputs from these structures are combined and processed using an MLP layer to predict ratings. This framework effectively addresses issues seen in traditional matrix factorization models, yielding notable performance enhancements and outperforming existing DL-based RS methods. Through extensive experiments on real-world datasets, the proposed model demonstrated superior performance in rating prediction compared to state-of-the-art techniques. These results suggest that DL methodologies are crucial for enhancing recommendation system performance. Our future direction includes integrating additional deep learning frameworks into recommendation systems to proactively boost recommendation quality. In addition, we plan to incorporate large language models into the recommender system domains.

Acknowledgement. The authors gratefully acknowledge the support of the ABSA Chair of Data Science and the Data Science for Social Impact (DSFSI) Lab at the University of Pretoria. This work was supported by UK International Development and the International Development Research Centre (IDRC), Ottawa, Canada, under the AI4D Africa Program. DSFSI also acknowledges gifts from NVIDIA, Google.org, OpenAI, and Meta.

References

- [1] Brent Smith and Greg Linden. “Two decades of recommender systems at Amazon. com”. In: *Ieee internet computing* 21.3 (2017), pp. 12–18.
- [2] Yong Zheng and David Xuejun Wang. “A survey of recommender systems with multi-objective optimization”. In: *Neurocomputing* 474 (2022), pp. 141–153.
- [3] Sen Li et al. “Embedding-based product retrieval in taobao search”. In: *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. 2021, pp. 3181–3189.
- [4] Aminu Da'u and Naomie Salim. “Recommendation system based on deep learning methods: a systematic review and new directions”. In: *Artificial Intelligence Review* 53.4 (2020), pp. 2709–2748.
- [5] Francesco Ricci, Lior Rokach, and Bracha Shapira. “Introduction to recommender systems handbook”. In: *Recommender systems handbook*. Springer, 2010, pp. 1–35.
- [6] Yehuda Koren. “Factor in the neighbors: Scalable and accurate collaborative filtering”. In: *ACM Transactions on Knowledge Discovery from Data (TKDD)* 4.1 (2010), pp. 1–24.
- [7] Shuai Zhang et al. “Deep learning based recommender system: A survey and new perspectives”. In: *ACM computing surveys (CSUR)* 52.1 (2019), pp. 1–38.
- [8] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. “Deep learning”. In: *nature* 521.7553 (2015), pp. 436–444.
- [9] Ayush Singhal, Pradeep Sinha, and Rakesh Pant. “Use of deep learning in modern recommendation system: A summary of recent works”. In: *arXiv preprint arXiv:1712.07525* (2017).
- [10] Yao Wu et al. “Collaborative denoising auto-encoders for top-n recommender systems”. In: *Proceedings of the ninth ACM international conference on web search and data mining*. 2016, pp. 153–162.
- [11] Xiangnan He et al. “Neural collaborative filtering”. In: *Proceedings of the 26th international conference on world wide web*. 2017, pp. 173–182.
- [12] Xiaoyuan Su and Taghi M Khoshgoftaar. “A survey of collaborative filtering techniques”. In: *Advances in artificial intelligence* 2009 (2009).
- [13] Mohammed Fadhel Aljunid et al. “A collaborative filtering recommender systems: Survey”. In: *Neurocomputing* 617 (2025), p. 128718.
- [14] Andriy Mnih and Russ R Salakhutdinov. “Probabilistic matrix factorization”. In: *Advances in neural information processing systems* 20 (2007).
- [15] Hong-Jian Xue et al. “Deep matrix factorization models for recommender systems.” In: *IJCAI*. Vol. 17. Melbourne, Australia. 2017, pp. 3203–3209.
- [16] Xiangnan He et al. “Fast matrix factorization for online recommendation with implicit feedback”. In: *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*. 2016, pp. 549–558.

- [17] Yehuda Koren, Robert Bell, and Chris Volinsky. “Matrix factorization techniques for recommender systems”. In: *Computer* 42.8 (2009), pp. 30–37.
- [18] Yiqi Gu et al. “Robust weighted SVD-type latent factor models for rating prediction”. In: *Expert Systems with Applications* 141 (2020), p. 112885.
- [19] Shijie Wang, Guiling Sun, and Yangyang Li. “SVD++ recommendation algorithm based on backtracking”. In: *Information* 11.7 (2020), p. 369.
- [20] Tongliang Liu and Dacheng Tao. “On the performance of manhattan non-negative matrix factorization”. In: *IEEE Transactions on Neural Networks and Learning Systems* 27.9 (2015), pp. 1851–1863.
- [21] Steffen Rendle. “Factorization machines with libfm”. In: *ACM Transactions on Intelligent Systems and Technology (TIST)* 3.3 (2012), pp. 1–22.
- [22] Shuai Zhang et al. “Deep learning for recommender systems”. In: *Recommender systems handbook*. Springer, 2021, pp. 173–210.
- [23] Huirong Guo et al. “DeepFM: a factorization-machine based neural network for CTR prediction”. In: *arXiv preprint arXiv:1703.04247* (2017).
- [24] Aanchal Mongia et al. “Deep latent factor model for collaborative filtering”. In: *Signal Processing* 169 (2020), p. 107366.
- [25] Taleb Alashkar et al. “Examples-rules guided deep neural network for makeup recommendation”. In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 31. 2017.
- [26] Gintare Karolina Dziugaite and Daniel M Roy. “Neural network matrix factorization”. In: *arXiv preprint arXiv:1511.06443* (2015).
- [27] Abebe Tamrat Tegene et al. “Deep Learning Based Matrix Factorization For Collaborative Filtering”. In: *2021 18th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP)*. IEEE. 2021, pp. 165–170.
- [28] Abebe Tegene et al. “Deep Learning and Embedding Based Latent Factor Model for Collaborative Recommender Systems”. In: *Applied Sciences* 13.2 (2023), p. 726.
- [29] Lei Zheng, Vahid Noroozi, and Philip S Yu. “Joint deep modeling of users and items using reviews for recommendation”. In: *Proceedings of the tenth ACM international conference on web search and data mining*. 2017, pp. 425–434.
- [30] Aminu Da'u, Naomie Salim, and Rabiu Idris. “An adaptive deep learning method for item recommendation system”. In: *Knowledge-Based Systems* 213 (2021), p. 106681.
- [31] George Cybenko. “Approximation by superpositions of a sigmoidal function”. In: *Mathematics of control, signals and systems* 2.4 (1989), pp. 303–314.
- [32] Binbin Hu et al. “Leveraging meta-path based context for top-n recommendation with a neural co-attention model”. In: *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*. 2018, pp. 1531–1540.

- [33] Yehuda Koren. “Factorization meets the neighborhood: a multifaceted collaborative filtering model”. In: *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. 2008, pp. 426–434.
- [34] Diederik P Kingma and Jimmy Ba. “Adam: A method for stochastic optimization”. In: *arXiv preprint arXiv:1412.6980* (2014).

Should AI Detect Social Stress? A Machine Learning Approach

Valentina Oelofse¹[0009-0009-2345-0497] and HMvE
Combrink^{1,2}[0000-0001-7741-3418]

¹ Department of Economics and Finance, University of the Free State, Bloemfontein,
South Africa

² Interdisciplinary Centre for Digital Futures, University of the Free State,
Bloemfontein, South Africa

Abstract. Social media has become a primary channel for communication and public discourse, yet its high-frequency and emotionally charged nature contributes to the spread of misinformation and elevated levels of social stress. Existing misinformation detection models do not account for the psychological and social toll of this environment. This study proposes a machine learning-based Social Stress Indicator (SSI) to detect and quantify stress signals in social media conversations. Using a synthetic dataset of annotated microblogs labelled by social scientists for social stress levels, four machine learning models-Logistic Regression, Random Forest, Naive Bayes, and K-Nearest Neighbour-were trained using TF-IDF embeddings. Sentiment features from VADER and RoBERTa were also integrated. Results showed consistently high performance across models, with accuracy above 99.4% and macro F1-scores exceeding 0.994. These findings demonstrate that machine learning models can reliably detect social stress from text data. However, the overperformance may reflect dataset homogeneity, necessitating further testing on real-world, diverse social media data to confirm generalisability. Future work should expand across platforms and contexts to validate this approach for stress-aware infodemic response.

Keywords: Infodemic Intelligence · Machine Learning · Social Stress · Social Media · Computational Infodemiology.

1 Introduction

The use of social media platforms has become a daily necessity for interpersonal engagements between co-workers, family and friends [?]. Social media serves as a platform where people can share information. Although benefits for the use of social media platforms exist, social media platforms also pose the risk of disseminating misinformation [?]. Furthermore, with the increase in social media, the dissemination of misinformation rapidly increases, which can give rise to the social stress of a community [?]. Misinformation spreading is not a new phenomenon, as misinformation has started to spread since the development of the printing press [?]. The major difference is that people now have easier access to social media platforms to spread misinformation on [?]. With the increase in misinformation, it necessitates establishing a social stress detection model that will ultimately be able to assess a given phrase and link a social stress label (low, medium, high) to the given phrase. Within social networks, misinformation spreads faster and with a wider reach than factual information [?]. With the faster sensation of views and clicks, the parties spreading the misinformation end up benefiting the most, while the communities viewing the misinformation experience higher levels of stress [?,?]. Some earlier studies described the knowledge gap between health information (which is seen as the evidence) and health misinformation (which is seen as what people think to be true), as an information epidemiology or as infodemiology [?]. Over time, the field of infodemiology research has grown and is recognised by the World Health Organisation as an emerging, novel scientific field [?]. This necessitates the importance of assessing how the infodemiology space within social media could affect the social stress of a community, which will ultimately help future research to understand the link between the economy and social media. This study aims to assist researchers in the field of infodemiology, with the intent to be able to use the social stress detection model in social media platforms like X, to be able to measure and detect high social stress. Four different machine learning algorithms were used, namely Random Forest, Logistic Regression, Naive Bayes and K-Nearest Neighbour. Each machine learning model was incorporated with Term Frequency-Inverse Document Frequency, which is a word frequency embedding. According to Zhau *et al.*, (2024), the addition of Term Frequency-Inverse Document Frequency improves the performance of the machine learning models [?,?]. A secondary dataset was used, which is based on the social stress work done by [?]. The dataset contains text-based data resembling social media conversations. These conversations were generated by a Large Language model. Social scientists labelled each phrase with the level of social stress that the phrase holds. The level of social stress was depended on the following five factors: anxiety, negativity, engagement level, help-seeking behaviour, and misinformation content. Although the Social Stress Indicator was found to be successful for detecting low levels of stress, medium and high levels were underestimated [?]. This establishes the need to further explore methods for the use of social stress detection. This study used the expertly labelled social stress dataset from the study by [?] and trained various machine learning models on the dataset. Furthermore, the performance

of the machine learning models was evaluated by using a confusion matrix as well as evaluation metrics such as precision, accuracy, recall and F1-score.

2 Literature Review

2.1 Infodemiology

Misinformation is false, incomplete or inaccurate information that is shared without the intention of inflicting harm [?,?]. Furthermore, this can be seen as information that is misleading or unintentionally spread without thinking that the information can inflict harm [?]. In contrast, disinformation is where false information is deliberately shared with the purpose of harming the public [?]. Disinformation is shared with the intent to deceive [?]. Not all types of disinformation have the same societal harm [?]. This emphasises the need to prioritise disinformation with the highest potential harm first [?]. An example is disinformation about politics, which can ultimately increase societal problems and erode norms [?]. Furthermore, disinformation is not only harmful to democratic institutions but also harmful to individuals' psychological well-being [?]. The rise in misinformation and disinformation can cause an absence of factual information, known as information voids or information vacuums [?]. When people see factual information infrequently, there is a lack of reliable information available amongst the pool of false information [?]. According to Palmer *et al.*, (2025), many people fall into the trap of the frequency bias, as people start to believe the information they see the most frequently [?]. This suggests that people will start believing misinformation just because there is an absence of credible and reliable information [?]. Infodemiology, also known as information epidemiology, has been widely used within health care systems to understand the distribution of health information and misinformation [?,?]. Infodemiology guides health care professionals and the public on the quality of health care information on social media platforms [?,?]. Infodemiology can be applied in fields such as public health monitoring as well as real-time disease surveillance [?]. The World Health Organisation acknowledged the term infodemic as the vast amount of accurate or false information that is spread [?]. Infodemiology has gained traction in recent years as a method to track disease outbreaks by analysing online data, which is also known as infoveillance [?]. Previous studies have demonstrated the use of infoveillance as a method to forecast Influenza outbreaks, map the sentiments of influenza vaccination, track cancer misinformation and Covid-19 misinformation [?,?,?]. Prior studies have also evolved disease surveillance into belief surveillance [?]. Belief surveillance is used in healthcare to measure the level of public belief based on health information posted on social media [?].

Although more and more health misinformation is being spread, not all users accept information equally. Many individuals and groups view information in different ways, which is known as information environments [?]. Information environments can largely be influenced by the amount and type of information present, how the information is arranged, and the social norms the individual or public believes in [?]. Belief systems are also based on the bandwagon effect,

where people are influenced by their social networks [?]. This means that people can endorse information just because everyone else endorses the information [?]. Belief systems can also be influenced by an individual's reference network, which is seen as people whom they care about [?]. In the next section, Social Stress will be discussed.

2.2 Social Stress

The psychological stress and harm due to societal pressures are more commonly known as social stress [?, ?, ?]. Social stress has become increasingly popular within the online interaction space, as social stress is more amplified and persistent online compared to face-to-face interactions [?]. Furthermore, social stress is also widely used in fields such as social psychology [?] and sociology [?, ?]. According to Wang *et al.*, (2019), mental stress is often a factor underestimated at an early stage, which evolves into severe health issues [?]. To address this, the use of social media platforms as a feasible platform to detect stress has been highlighted [?]. Moreover, a Factor Graph Model combined with a Convolutional Neural Network was used on Twitter (X) data to measure stress [?]. Additionally, machine learning was used for sentiment analysis to explore stress within social media [?]. Kumari *et al.*, (2022), found that the best performing machine learning models based on accuracy were Random Forest, Support Vector Machine and Logistic regression that incorporated Term Frequency-Inverse Document Frequency [?]. Social media is a platform where the public can express their opinions and retrieve and disseminate information [?]. With the increased use of social media, it necessitates a need to measure the public's perceptions and sentiment in real time [?].

Social media analytics can be used to measure the real-time social stress of societal pressures by making use of indicators. Various indicators have been used to measure stress and anxiety, namely Sentiment Analysis and Topic Modelling [?]. While information-seeking behavior is more commonly used to measure trends, information-seeking behavior can also be used to capture real-time social stress on social media platforms [?]. Sentiment analysis is a field of study that analyses people's opinions, emotions and appraisals towards various topics, in text-based content [?, ?, ?]. Sentiment analysis can also be referred to as subjectivity analysis, appraisal extraction as well as opinion mining [?]. Sentiment analysis is seen as the process of using text as the input and extracting valuable information about the sentiment as the output [?]. Examples of the use of sentiment analysis include emotion recognition applications that detect emotions through facial expressions, as well as performing credit ratings and entity reputation evaluations. Sentiment classification is a natural language processing task that is used to determine the sentiment or the emotional tone by classifying the text as either positive, negative or neutral [?, ?]. Various authors made use of the TextBlob Library in machine learning to be able to evaluate the emotions

in the text-based content [?]. Sentiment Analysis can be defined as (Eq. 1) [?]:

$$SA = - \sum_{I=0}^N \left(\frac{x_1 + x_2 + x_3 + \dots + x_n}{N} \right), \quad (1)$$

where SA is seen as the sentiment and N is the sample size of the total number of microblogs measured, and $x_1, x_2, x_3, \dots, x_n$ represents the sentiment per individual microblog for a specific time stamp [?].

Topic modelling is seen as a text mining technique to identify themes or topics present in a large corpus of text-based data [?]. Topic modelling is useful within the social media analytics space as it is able to identify and track relevant topics of concern [?]. According to Ji *et al.*, (2025), the Latent Dirichlet Allocation (LDA) unsupervised machine learning algorithm has become a popular method among researchers to perform topic modelling [?]. LDA assigns a topic to each word within the corpus, while using probabilistic methods to determine the frequency the word has appeared within the topics as well as the distribution of the topics within the corpus [?]. Topic modelling can be defined through the coherence score as (Eq. 2) [?]:

$$C(t_k) = \sum_{i=1}^n \sum_{j=i+1}^n \log \left(\frac{P(W_i, W_j) + \epsilon}{P(W_i) \cdot P(W_j)} \right), \quad (2)$$

where $P(W_i, W_j)$ denotes the probability of words W_i and W_j co-occurring and $P(W_i)$ and $P(W_j)$ represent the probabilities of W_i and W_j occurring respectively, and ϵ is seen as a smoothing term [?]. When individuals look for information online to gain knowledge in different topics, it is known as information-seeking behaviour [?]. Information-seeking behavior can provide useful psychological and socioeconomic insight [?]. Information-seeking behavior as an indicator can be seen as the number of people who searched for a specific topic over a certain period on a specific platform [?]. Google searches have been seen as a reliable platform to assess online interactions, which can be tracked with Google Trends [?]. Google Trends provides real-time data about a keywords search volume within a selected timeframe and geography [?]. Information-seeking behaviour can also predict behavioural risks [?]. The equation for information-seeking behaviour can be defined as (Eq. 3) [?]:

$$ISB = \frac{\sum_{i=0}^N \beta}{\sum_{i=0}^N \bar{X}}, \quad (3)$$

where ISB is the information-seeking behaviour and β is the frequency of the same topic at a particular point in time, denoted as a percentage, which is specific to a series [?].

The Social Stress Indicator for social media was developed as a computational tool to quantify and measure social stress in real time, using sentiment analysis, subjectivity, and information-seeking behaviour [?]. The equation for the Social Stress Indicator (SSI) was defined as (Eq. 5) [?]:

$$SSI = \frac{-SA + (1 - SUB) + ISB}{3}, \quad (4)$$

where, SA refers to the sentiment, specifically focusing on the negative sentiment, SUB as the subjectivity and ISB as the information-seeking behaviour. Eq. 5 currently uses equal weightings for each component of the Social Stress Indicator. Moreover, the Social Stress Indicator was shown to capture real-time low levels of stress successfully, while the medium and high levels of stress were severely underestimated [?]. This study used the expertly labelled social stress dataset introduced in prior work and trained various machine learning models on the dataset [?].

3 Methods

This study followed a structured methodological framework comprising the design of the Social Stress Indicator (SSI), the creation of a synthetic dataset, the application of natural language preprocessing and feature engineering techniques, and the training and evaluation of four machine learning models. The methodology was designed to ensure that each stage of the process could be replicated and independently validated by other researchers.

The core of this approach is the Social Stress Indicator (SSI), which operationalises social stress as a composite measure of three text-derived indicators namely sentiment, subjectivity, and information-seeking behaviour. As defined in Eq. 5, the SSI is expressed as:

$$SSI = \frac{-SA + (1 - SUB) + ISB}{3}, \quad (5)$$

where *SA* denotes the sentiment score (focused on negative sentiment), *SUB* refers to the subjectivity score, and *ISB* represents the information-seeking behaviour measured as the normalised frequency of searches or interactions for a given topic over time. Each component was scaled to the [0, 1] interval, and equal weighting was applied to reflect their combined influence on social stress. The negative sign for *SA* captures the effect of negative sentiment on the stress level, while $(1 - SUB)$ inversely relates subjectivity to stress, acknowledging that more objective content may attenuate perceived stress. The SSI was computed for each microblog entry to produce a continuous score, which was then discretised into three classes (Low, Medium, High) based on thresholds co-developed with domain experts [?].

The dataset used to develop and evaluate the SSI was synthetically generated to emulate the structure and content of real-world social media posts. Prompt engineering with a large language model was employed to produce microblog-like texts covering diverse themes such as health misinformation, financial insecurity, family issues, and public crises. This ensured coverage of multiple stress-inducing contexts. Each generated post was manually annotated by three independent raters with expertise in social sciences, using a rubric derived from psychological and sociological stress frameworks. The rubric assessed five attributes: anxiety, negativity, engagement level, help-seeking behaviour, and misinformation content [?]. Each post received a stress classification label corresponding to Low, Medium, or High stress. Inter-rater agreement achieved a Cohens κ score of 0.82, indicating strong reliability. The final dataset comprised 3,000 posts, equally balanced across the three stress levels.

Prior to feature extraction, the text was preprocessed to ensure consistency and to remove noise. All text was converted to lowercase, punctuation was stripped, and stopwords were removed using the Natural Language Toolkit (NLTK) [?]. Tokenisation and lemmatisation were subsequently applied to standardise word forms [?]. Each microblog was then represented by a high-dimensional feature vector constructed from a combination of lexical and engineered features. Term FrequencyInverse Document Frequency (TFIDF) embeddings were generated using scikit-learns `TfidfVectorizer` with a maximum of 5,000 features, a bigram range of (1, 2), and a minimum document frequency threshold of five. Sentiment features were extracted from two complementary sources: the VADER sentiment analysis tool, which provides lexicon-based compound sentiment scores in the range $[-1, 1]$, and the RoBERTa transformer model for contextual sentiment classification using the `cardiffnlp/twitter-roberta-base-sentiment` model. Subjectivity scores were derived from TextBlob and normalised to $[0, 1]$. Information-seeking behaviour (ISB) was calculated as the normalised frequency of topic-related searches over time, following the approach described in [?]. These features were concatenated to form a single feature matrix of dimension $(N, 5003)$, comprising 5,000 TFIDF features and three engineered features (VADER, RoBERTa, and ISB).

Four supervised machine learning models were selected to evaluate the SSI, namely Logistic Regression, Random Forest, Multinomial Naïve Bayes, and K-Nearest Neighbour (KNN). Logistic Regression was implemented with an L2 penalty and the `lbfgs` solver in multinomial mode. Random Forest was configured with 200 trees and unrestricted depth using the Gini impurity criterion. Multinomial Naïve Bayes used Laplace smoothing with $\alpha = 1.0$, while KNN used five neighbours and Euclidean distance as the metric. Hyperparameters for each model were tuned using a five-fold grid search to optimise performance. The dataset was split into training and testing subsets using an 80/20 stratified split to preserve the balance of stress levels across subsets.

Model performance was evaluated using a comprehensive set of metrics, including accuracy, macro-averaged precision, recall, and F1-score, as well as weighted averages for each metric. Confusion matrices were generated for per-

class error analysis to assess misclassification patterns. All analyses were conducted in Python 3.10 using scikit-learn 1.3, pandas 2.0, NLTK 3.8, and Hugging Face Transformers 4.30 in Google Colab Pro with an NVIDIA T4 GPU and 12 GB RAM. This environment ensured reproducibility and scalability of the experiments for a reproducible baseline for stress-aware infodemic analytics.

4 Results and Discussion

A subset of annotated microblogs is presented in Table 1 to illustrate the scoring mechanism used in the social stress analysis. From the experiments, the model can distinguish between varying degrees of emotional content. For instance, positive or neutral expressions such as "Life is smooth and simple." received a low score of 0.20, indicating minimal psychosocial strain. Conversely, posts conveying acute psychological burden, such as "This is too much the pressure of expectations.", yielded scores exceeding 0.90, reflecting high-intensity stress signals. Of particular interest are the intermediate cases, such as "Staying hopeful about family issues." (0.32), where the sentiment is cautiously optimistic yet still linked to an underlying stressor. This is particularly relevant for domains such as mental health monitoring, crisis intervention, and infodemic management, where understanding the spectrum of stress expressions can inform timely and proportionate responses.

Table 1. Subset of annotated microblogs

Microblog	Social Stress Score
Life is smooth and simple.	0.20
This is too much the pressure of expectations.	0.92
Staying hopeful about family issues.	0.32
Losing sleep over family issues.	0.88

Table 2 illustrates the tokenized representation of microblog content, paired with their normalized SSS and corresponding stress classification labels. This representation highlights the transition from raw narrative text to structured analytical units suitable for computational modelling. Stress labels are assigned according to pre-defined thresholds with 0 for low or negligible stress, 1 for moderate stress, and 2 for high stress intensity. For example, the token sequence [life, smooth, simple] reflects a low-stress state (0.20, label 0), while [losing, sleep, family, issues] reveals a high-stress profile (0.88, label 2) associated with family-related anxieties. Tokens such as [staying, hopeful, family, issues], with a moderate score of 0.32 (label 1), demonstrate that models can capture nuanced emotional states where positive sentiment coexists with underlying social stressors, but this is specific to the data.

Table 2. Microblog Tokens with SSS Normalized Scores and Stress Labels

Microblog Tokens	SSS Normalized Stress Label		
[life, smooth, simple]	0.20	0	
[staying, hopeful, family, issues]	0.32	1	
[much, pressure, expectations]	0.92	2	
[losing, sleep, family, issues]	0.88	2	

Table 3 presents tokenized microblogs enriched with both stress labels and sentiment polarity scores generated from two distinct natural language processing frameworks, such as VADER and RoBERTa. The inclusion of sentiment analysis alongside stress classification enables a multidimensional interpretation of each microblog, capturing not only the presence of stress but also its affective orientation. For instance, the token sequence [life smooth simple], labelled as low stress (0), is consistently identified as predominantly positive by both models, with RoBERTa attributing a notably high positive score (0.79) compared to VADERs binary-neutral output (1.00 in the neutral category). In contrast, a high-stress sequence such as [losing sleep family issues] received overwhelmingly negative sentiment scores from both frameworks, particularly RoBERTa (0.87), highlighting its heightened sensitivity to emotionally charged language. Intermediate stress cases, such as [staying hopeful family issues] (label 1), exhibit mixed sentiment patterns, where positive sentiment persists despite the presence of stress-related terms, demonstrating the models ability to detect nuanced emotional states.

Table 3. Microblog Tokens with Stress Labels and Grouped Sentiment Scores

Microblog Tokens	Stress Label	VADER			RoBERTa		
		Neg	Neu	Pos	Neg	Neu	Pos
[life smooth simple]	0	0.00	1.00	0.00	0.01	0.20	0.79
[staying hopeful family issues]	1	0.00	0.55	0.45	0.01	0.32	0.67
[much pressure expectations]	2	0.24	0.76	0.00	0.75	0.23	0.02
[losing sleep family issues]	2	0.39	0.61	0.00	0.87	0.12	0.01

Across metrics, Logistic Regression, Random Forest, and Naive Bayes exhibit near-identical, near-ceiling results (Accuracy ≈ 0.9965 , Macro/Weighted $F_1 \in [0.9963, 0.9965]$), while KNN trails marginally (Accuracy = 0.9947; Macro $F_1 = 0.9947$), Table 4. The close correspondence between macro- and weighted-averaged scores suggests limited class imbalance in the evaluation set and indicates that gains are not driven solely by majority classes. The parity among

three distinct model families (linear, probabilistic, and ensemble) implies that the TF-IDF feature space already affords highly separable decision boundaries, rendering additional model complexity comparatively unimportant. While these results demonstrate excellent in-sample generalisation under the current protocol, such uniformly high scores warrant caution: they may reflect dataset homogeneity, lexical cues that trivially correlate with labels, or residual leakage between train and test partitions. We therefore recommend complementary diagnostics (per-class confusion matrices, calibrated probability assessment, and error analysis of false positives/negatives) and validation checks (stratified cross-domain validation, temporal splits, and perturbation tests) to confirm that the observed performance translates beyond the present sample and to identify any over-reliance on spurious lexical markers.

Table 4. Classification performance comparison across models using TF-IDF features.

Metric	Logistic Regression	Random Forest	Naive Bayes	KNN
Accuracy	0.9965	0.9965	0.9965	0.9947
Macro Avg Precision	0.9969	0.9969	0.9969	0.9954
Macro Avg Recall	0.9957	0.9957	0.9957	0.9940
Macro Avg F1-score	0.9963	0.9963	0.9963	0.9947
Weighted Avg Precision	0.9965	0.9965	0.9965	0.9948
Weighted Avg Recall	0.9965	0.9965	0.9965	0.9947
Weighted Avg F1-score	0.9965	0.9965	0.9965	0.9947

Table 5 presents the classification performance of four machine learning models trained with two different composite feature sets: *TF-IDF + RoBERTa* and *TF-IDF + VADER*. Across all metrics and both feature configurations, performance is uniformly high, with accuracy values exceeding 0.994 and negligible variation between models.

Table 5. Performance of ML models with different feature sets.

Metric	TF-IDF + RoBERTa					TF-IDF + VADER				
	Logistic Regression	Random Forest	Naive Bayes	KNN		Logistic Regression	Random Forest	Naive Bayes	KNN	
Accuracy	0.9953	0.9953	0.9953	0.9941	0.9953	0.9953	0.9953	0.9953	0.9941	
Macro Avg Precision	0.9957	0.9957	0.9957	0.9947	0.9957	0.9957	0.9957	0.9957	0.9947	
Macro Avg Recall	0.9947	0.9947	0.9947	0.9936	0.9947	0.9947	0.9947	0.9947	0.9936	
Macro Avg F1-score	0.9952	0.9952	0.9952	0.9941	0.9952	0.9952	0.9952	0.9952	0.9941	
Weighted Avg Precision	0.9953	0.9953	0.9953	0.9942	0.9953	0.9953	0.9953	0.9953	0.9942	
Weighted Avg Recall	0.9953	0.9953	0.9953	0.9941	0.9953	0.9953	0.9953	0.9953	0.9941	
Weighted Avg F1-score	0.9953	0.9953	0.9953	0.9941	0.9953	0.9953	0.9953	0.9953	0.9941	

Logistic Regression, Random Forest, and Naive Bayes exhibit identical scores for each metric within a given feature set, while KNN lags marginally, particularly in accuracy and macro-averaged metrics, by approximately 0.001–0.002.

Notably, the integration of sentiment analysis features from RoBERTa and VADER does not appear to yield differential performance improvements, as the two composite feature sets produce identical values to four decimal places for each model. This suggests that, within the constraints of the present dataset, sentiment-derived features may be redundant when combined with high-dimensional lexical representations from TF-IDF. The homogeneity of results across diverse model architectures further implies that the task is linearly separable in the transformed feature space, and that model choice exerts minimal influence on predictive accuracy. While these findings reflect strong in-sample performance, they also raise concerns regarding potential overfitting, data leakage, or limited variability in the evaluation set. To ensure external validity, future work should investigate the stability of these results under domain-shift conditions, perform stratified temporal validation, and conduct feature importance analyses to assess the relative contribution of sentiment-based features in broader, more heterogeneous corpora. By training on more data, the models can learn how to perform better with unknown samples. Given that this was a relatively small sample size, it is suggested that more data be introduced for social stress classification, as the ML approach to a SSS can be applied. While this study demonstrates promising results using machine learning models trained on a synthetically generated dataset, it is essential to acknowledge the implications of using such data on performance evaluation. The synthetic microblogs were generated using a large language model and annotated by domain experts, allowing control over topic balance and stress-level representation. However, synthetic data may not fully capture the complexity, noise, ambiguity, or adversarial phrasing typical of real-world social media content. This could result in inflated performance metrics due to reduced linguistic variability and more distinct decision boundaries between classes. The near-ceiling accuracy and F1-scores observed across all models suggest that the dataset may be overly homogeneous or artificially well-separated, which limits the generalisability of these results.

Real-world social media posts are often contextually nuanced, highly informal, and influenced by sarcasm, idioms, or multimedia context, which synthetic data might fail to emulate. Additionally, platform-specific trends, evolving language, and demographic markers embedded in real conversations are critical components for building robust social stress detection systems. Due to privacy concerns, ethical data collection restrictions, and annotation challenges related to sensitive psychological content, real-world labelled datasets for social stress remain scarce. Therefore, synthetic data was employed as an initial proof of concept to establish baseline model efficacy while mitigating topic and label bias. Therefore, future work should involve external validation on real-world datasets and explore domain adaptation methods to transition from synthetic to naturalistic data environments, ensuring that these models are ethically deployable in high-stakes settings such as mental health surveillance or infodemic early warning systems.

5 Conclusion

This study was set out to evaluate the capacity of multiple machine learning models, namely Logistic Regression, Random Forest, Naive Bayes, and KNN, to predict the SSS. Results from this study demonstrate that all evaluated machine learning models, when applied with either TF-IDF + RoBERTa or TF-IDF + VADER feature sets, achieved near-identical and exceptionally high performance metrics across accuracy, precision, recall, and F1-scores. Specifically, Logistic Regression, Random Forest, and Naive Bayes consistently recorded accuracy values of 0.9953, while KNN marginally trailed with an accuracy of 0.9941. The uniformity of macro and weighted averages across models and feature combinations indicates that the classification task posed minimal challenge to the algorithms, suggesting a potential ceiling effect caused by the datasets structure or class distribution. While such high scores may initially appear to reflect excellent model generalisation, they also raise concerns about the underlying data complexity and diversity. In tasks with limited variance or highly separable classes, inflated metrics can obscure a models ability to generalise to unseen, real-world scenarios. The negligible difference between the RoBERTa and VADER sentiment integration further suggests that the choice of sentiment extraction method had little influence on predictive performance within this dataset. This convergence in outcomes underscores the importance of evaluating not only statistical performance but also the representativeness and challenge level of the training data. Without a dataset that captures the variability and subtlety of the problem space, even state-of-the-art models may deliver impressive yet potentially misleading results, thereby limiting their applicability in practical, high-stakes decision-making contexts. Therefore, should we trust AI to detect social stress? In the context of this study, more experimentation is required to implement such models in a real-world scenario, but based on this experiment, if the models are trained correctly, there is promise and potential for the implementation of such models.

Cross-lingual transfer of multilingual models on low resource African Languages

Harish Thangaraj¹ , Ananya Chenat¹, Jaskaran Singh Walia¹ , and Vukosi Marivate^{2,3} 

¹ Vellore Institute of Technology, India

² Data Science for Social Impact, University of Pretoria, South Africa

³ Lelapa AI

Abstract. Cross-lingual transfer learning is crucial for developing NLP technologies for low-resource African languages, yet the optimal modeling strategy remains an open question. This paper presents a benchmark for cross-lingual transfer from Kinyarwanda to Kirundi, two closely related Bantu languages. We evaluate the performance of traditional monolingual architectures (BiGRU, CNN, Char-CNN) against three distinct multilingual transformer models: the global mBERT, the Africa-centric AfriBERT, and the language-family-specific BantuBERTa. Our evaluation covers both zero-shot and fine-tuned news classification, and critically, we measure the degree of catastrophic forgetting on the source language after fine-tuning. Our results demonstrate that the regionally-focused AfriBERT achieves the highest cross-lingual accuracy (88.3%) after fine-tuning. Furthermore, we find that large-scale pre-training is essential for robustness; mBERT and AfriBERT exhibit minimal forgetting, while BantuBERTa and the traditional models suffer a severe performance degradation. This study highlights the importance of regionally focused multilingual models for transfer between related African languages and establishes catastrophic forgetting as a critical evaluation metric for such tasks.

Keywords: Cross-lingual Transfer · Low-resource Language Modelling · Multilingual Models · Catastrophic Forgetting · Kinyarwanda and Kirundi · News Classification

1 Introduction

The development of robust Natural Language Processing (NLP) technologies for the world’s diverse languages remains a significant challenge, particularly for the over 2000 languages spoken across Africa. Due to data scarcity, cross-lingual transfer learning—where learnings from a relatively higher-resource language is adapted for a lower-resource one—has become an essential strategy. The effectiveness of this transfer, however, depends critically on the choice of model architecture and pre-training regimen.

A central debate revolves around the optimal approach for transfer. On one hand, traditional monolingual models (e.g., CNN, BiGRU) can be trained to capture the specific nuances of a source language. On the other, large multilingual transformer models like mBERT [4] offer powerful, pre-trained representations learned from a global

corpus. More recently, regionally-focused models such as AfriBERT [14] and BantuBERTa [16,17] have emerged, promising more specialised transfer capabilities by pre-training on curated sets of related languages. Yet, there is a lack of comprehensive benchmarks that compare these different strategies head-to-head in a realistic African language context, particularly regarding the crucial trade-off between transfer performance and the risk of catastrophic forgetting—where a model loses its proficiency in the original source language after fine-tuning.

This paper addresses this gap by conducting a rigorous benchmarking study of cross-lingual transfer from Kinyarwanda to Kirundi, two closely related Bantu languages. Extending the work of [13], we compare traditional neural architectures against global, regional, and language-family-specific multilingual transformer models. Our core research question is: Which transfer learning strategy yields the best performance on the target language while retaining performance on the source language?

Our main contributions are:

1. A comprehensive benchmark of six models (mBERT, AfriBERT, BantuBERTa, Bi-GRU, CNN, Char-CNN) for Kinyarwanda-to-Kirundi news classification, evaluated in both zero-shot and fine-tuned settings.
2. An empirical analysis of catastrophic forgetting, quantifying how different model architectures are affected by fine-tuning on a low-resource target language.
3. A clear demonstration of the superiority of regionally-focused multilingual models (AfriBERT) for this task, providing practical guidance for researchers and practitioners working on related African languages.

2 Related Work

Our research is situated at the intersection of cross-lingual transfer learning, model architecture, and Natural Language Processing for low-resource African languages.

2.1 Cross-Lingual Transfer for Low-Resource Languages

Addressing data scarcity in low-resource languages (LRLs) is a central challenge in modern NLP [10]. Cross-lingual transfer learning has emerged as a dominant strategy, where learnings from a higher-resource source language is transferred to a lower-resource target language. Early approaches demonstrated the viability of transfer using techniques like annotation projection and direct transfer with LSTM architectures, showing that even single-source transfer from a distant language like English can improve performance [20]. More recent work has focused on improving transfer efficiency through data augmentation [19], back-translation [7], and sophisticated adapter-based architectures that leverage multi-source information [18]. Our work builds on the "direct transfer" approach, focusing on a scenario where the source and target languages are closely related.

2.2 Paradigms in Transfer: Monolingual vs. Multilingual Models

A key debate in cross-lingual transfer concerns the optimal model architecture. One paradigm focuses on transferring learnings from powerful **monolingual models**. Studies have shown that representations from a single-language BERT can be effectively transferred to other languages, preserving both syntactic and semantic knowledge [6]. Monolingual models fine-tuned on specific dialects, such as DarijaBERT for Moroccan Arabic, have even been shown to outperform larger multilingual models on specific tasks, highlighting their ability to capture fine-grained linguistic nuances [2].

The alternative paradigm leverages large **multilingual models** like mBERT [4] and XLM-R [3], which learn a shared representation space across many languages during pre-training. These models have demonstrated strong zero-shot and few-shot cross-lingual capabilities [21]. Their effectiveness often depends on aligning latent representations between languages, for which methods like unsupervised machine translation have been explored [5]. However, the performance of these global models can be surpassed by regionally-focused multilingual models, such as ARBERT and MARBERT for Arabic, which achieve state-of-the-art results by pre-training on a curated set of related languages and dialects [1]. Our paper directly contributes to this debate by empirically comparing the cross-lingual performance of both monolingual and various multilingual models in a resource-constrained setting.

2.3 Advancements in NLP for African Languages

There is a growing body of work dedicated to developing NLP resources and models specifically for African languages. Research has shown that multilingual training incorporating related languages significantly enhances performance for languages like isiZulu and Sepedi [11], and multilingual neural machine translation models consistently outperform single-pair models for African languages [9].

This has led to the development of Africa-centric language models. Studies have demonstrated that multilingual models pre-trained on African languages outperform monolingual baselines in transfer tasks [15] [12]. This motivates our use of **AfriBERT** [14], which is pre-trained on 11 African languages, and **BantuBERTa** [16,17], which focuses specifically on the Bantu language family. Our work serves as a direct extension of [13], which introduced the KINNEWS and KIRNEWS datasets and provided initial benchmarks using traditional architectures like BiGRU and CNN. By including and benchmarking modern, Africa-centric transformer models against these baselines, we provide an updated and more comprehensive analysis of cross-lingual transfer between Kinyarwanda and Kirundi.

3 Experiments

We conduct a series of experiments to benchmark cross-lingual transfer from Kinyarwanda to Kirundi. Our methodology is organised into four parts: the datasets used, the models evaluated, the experimental training protocol, and the evaluation metrics.

3.1 Datasets and Preprocessing

The study utilises the KINNEWS (Kinyarwanda) and KIRNEWS (Kirundi) datasets introduced by [13]. These two languages belong to the Rwanda-Rundi dialect continuum and share significant lexical and grammatical similarities, making them ideal candidates for a cross-lingual transfer study.

The raw data was preprocessed by removing duplicate articles and irrelevant metadata. For a unified text representation, the ‘title’ and ‘content’ of each news article were concatenated. The datasets were filtered to include only the 12 semantic categories common to both languages. A summary of the resulting datasets is provided in Table 1.

Table 1: Statistics of the preprocessed KINNEWS and KIRNEWS datasets used in this study.

Language	Dataset	Articles	Classes	Train/Test Split
Kinyarwanda	KINNEWS	21,268	12	17,014 / 4,254
Kirundi	KIRNEWS	4,612	12	3,690 / 922

3.2 Models

We evaluate two classes of models to compare their cross-lingual transfer capabilities.

Multilingual Transformer Models. We use three pre-trained transformer-based models known for their multilingual capabilities:

- **mBERT** [4]: A multilingual BERT model pre-trained on Wikipedia text from 104 languages, serving as a strong general-purpose cross-lingual baseline.
- **AfriBERT** [14]: A model pre-trained on a curated corpus of 11 African languages, including Kinyarwanda and Kirundi, designed to better capture regional linguistic nuances.
- **BantuBERTa** [16,17]: A model pre-trained specifically on Bantu languages, intended to leverage the typological similarities within this language family.

Traditional Neural Models. As monolingual baselines, we adapt three standard neural architectures from [13], which rely on language-specific embeddings:

- **CNN**: A Convolutional Neural Network for text classification as proposed by [8].
- **Char-CNN**: A character-level CNN, capable of capturing sub-word morphological features.
- **BiGRU**: A Bidirectional Gated Recurrent Unit network, which processes text sequentially to capture contextual information.

For these models, we trained a Word2Vec model (skip-gram, vector size 50, window size 5) on the Kinyarwanda training corpus to generate input word embeddings.

3.3 Experimental Design and Training

Our experimental protocol involves a three-stage process to comprehensively evaluate cross-lingual transfer and catastrophic forgetting.

1. **Source Language Training:** All models are first trained on the Kinyarwanda training set for the 12-class news classification task which was split into 90% for training and 10% for validation..
2. **Target Language Evaluation:** The Kinyarwanda-trained models are then evaluated on the Kirundi test set in two distinct settings:
 - *Zero-Shot Transfer*: The model is evaluated directly on the Kirundi test set without any further training.
 - *Fine-Tuned Transfer*: The model is fine-tuned on the Kirundi training set before being evaluated on the Kirundi test set.
3. **Forgetting Evaluation:** After fine-tuning on Kirundi, the models are re-evaluated on the Kinyarwanda test set to measure the degradation in source-language performance.

The transformer models were trained using the hyperparameters detailed in Table 2 and the Traditional models were adopted from [13]

Table 2: Hyperparameters for fine-tuning the transformer models.

Parameter	Value
Number of Labels	12
Input Sequence Length	128
Truncation	True
Padding	True
Device	MPS
Number of Training Epochs:	
mBERT	8
AfriBERT	25
BantuBERTa	8
Per-Device Train/Eval Batch Size	32
Warmup Steps	500
Weight Decay	0.01
Logging Steps	10
Load Best Model at End	True
Metric for Best Model	f1
Evaluation Strategy	steps

All experiments were repeated five times with different random seeds, and we report the mean of accuracy and F1-score with least standard deviations.

3.4 Evaluation Metrics

We use three metrics to assess model performance across different stages of the experiment:

- **Accuracy:** The proportion of correctly classified examples.
- **F1-Score:** The macro-averaged F1-score, which is the harmonic mean of precision and recall, providing a balanced measure for multi-class classification.
- **Forgetting:** The percentage decrease in performance on the original (Kinyarwanda) task after the model has been fine-tuned on the new (Kirundi) task. It is calculated as:

$$\text{Forgetting (\%)} = \left(\frac{\text{Acc}_{\text{initial}} - \text{Acc}_{\text{final}}}{\text{Acc}_{\text{initial}}} \right) \times 100 \quad (1)$$

where $\text{Acc}_{\text{initial}}$ is the accuracy on Kinyarwanda before fine-tuning on Kirundi, and $\text{Acc}_{\text{final}}$ is the accuracy after.

4 Results

This section presents the empirical results of our cross-lingual transfer experiments. We report on three key aspects: (1) the baseline performance of models on the source language (Kinyarwanda), (2) the cross-lingual performance on the target language (Kirundi) before and after fine-tuning, and (3) the extent of catastrophic forgetting on the source language after fine-tuning.

4.1 Cross-Lingual Transfer Performance

The core results for the Kinyarwanda-to-Kirundi transfer task are presented in Table 3. In the zero-shot setting (Before FT), the multilingual models demonstrated a clear advantage, with BantuBERTa achieving the highest accuracy (74.5%). In contrast, the traditional models performed near random chance. After fine-tuning on Kirundi data (After FT), all models showed substantial gains. AfriBERT emerged as the top-performing model in terms of accuracy, reaching **88.3%**. Notably, the BiGRU model achieved the highest F1-score at **87.9%**, highlighting its strength as a traditional baseline.

Table 3: Metrics describing cross-lingual testing on Kirundi

Model	Accuracy before FT	F1 before FT	Accuracy after FT	F1 after FT
mBERT	0.5872	0.5917	0.8462	0.8422
AfriBERT	0.7421	0.7474	0.8830	0.8787
BantuBERTa	0.7454	0.7375	0.8657	0.8606
BiGRU	0.2404	0.2300	0.8332	0.8790
CNN	0.2190	0.2320	0.5913	0.5732
Char-CNN	0.1916	0.1621	0.4879	0.4764

4.2 Monolingual Performance and Catastrophic Forgetting

Table 4 details model performance on the source Kinyarwanda dataset. Before fine-tuning (Table 4a), the traditional models BiGRU and CNN achieved the highest monolingual accuracy at 88.5% and 87.4%, respectively. Table 4b shows the performance on Kinyarwanda after the models were fine-tuned on Kirundi. The large multilingual models, mBERT and AfriBERT, proved highly resilient, retaining most of their original performance with forgetting rates of just **3.0%** and 5.1%. In stark contrast, BantuBERTa and all traditional models suffered from severe catastrophic forgetting, with performance drops exceeding 70%. Figure 1 visually summarises this degradation.

Table 4: Performance on the Kinyarwanda test set. (a) Initial monolingual performance before any cross-lingual fine-tuning. (b) Performance after models were fine-tuned on Kirundi, with the calculated percentage of forgetting. Best results are in **bold**.

(a) Performance before fine-tuning.			(b) Performance after fine-tuning on Kirundi.		
Model	Accuracy	F1 score	Model	Accuracy	Forget %
mBERT	0.7884	0.7747	mBERT	0.7645	3.03
AfriBERT	0.8498	0.8447	AfriBERT	0.8061	5.14
BantuBERTa	0.8601	0.8555	BantuBERTa	0.2172	74.00
BiGRU	0.8851	0.8434	BiGRU	0.2329	73.68
CNN	0.8740	0.8660	CNN	0.2207	74.86
Char-CNN	0.6930	0.6823	Char-CNN	0.1968	71.50

Figure 1 portrays a graphical representation of forgetting and improvement after fine-tuning, for Kinyarwanda.

5 Discussion

Our results offer several key insights into cross-lingual transfer for low-resource African languages. While baseline models achieve the highest overall F1 performance before fine-tuning, multilingual models demonstrate competitive results and exhibit clear advantages in retaining source-language performance after fine-tuning. This highlights the strength of pre-trained representations as a foundation that monolingual models trained from scratch on small datasets cannot match.

The standout performance of AfriBERT (88.3% accuracy) over mBERT (84.6%) is noteworthy. AfriBERT is trained on a focused set of 11 African languages, which allows it to learn representations that capture shared lexical, morphological, and syntactic patterns within this group. In contrast, mBERT is trained on 104 languages, and the uneven distribution of corpora across languages can reduce its ability to model low-resource languages effectively. These factors likely contribute to AfriBERT’s superior

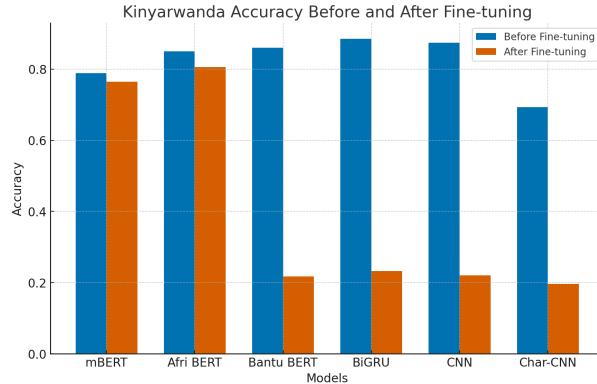


Fig. 1: Performance degradation on Kinyarwanda after fine-tuning, illustrating the forgetting gap between large multilingual models and others.

performance, indicating that limiting pretraining to a smaller, linguistically related set of languages can improve cross-lingual transfer, especially for Bantu languages.

critical finding is the large difference in catastrophic forgetting across models. AfriBERT and mBERT show low forgetting (5.1% and 3.0%), whereas BantuBERTa and the traditional models experience much higher forgetting (74.0% and near-total loss). We hypothesise that this is influenced by the scale and diversity of pre-training and also by tokenizer coverage: models with better tokenization for Kinyarwanda and Kirundi are likely to retain knowledge more effectively. In contrast, models with smaller pre-training or less suitable tokenizers are more prone to forgetting. This suggests that both pre-training scale and tokenization quality contribute to the observed differences.

Finally, the results underscore the value of linguistic similarity. The high degree of mutual intelligibility between Kinyarwanda and Kirundi provides an ideal scenario for transfer learning. Our work empirically demonstrates that modern transformer architectures are highly effective at leveraging this lexical and structural overlap, even with limited fine-tuning data.

6 Conclusion

This study benchmarked the cross-lingual transfer capabilities of multilingual and traditional neural models on a Kinyarwanda-to-Kirundi news classification task. We demonstrated that pre-trained multilingual models, particularly the regionally-focused AfriBERT, significantly outperform traditional architectures in transfer accuracy. Our analysis also quantified the critical challenge of catastrophic forgetting, revealing the superior resilience of models with larger pre-training corpora. The core contribution of this work is an empirical comparison that provides clear guidance on model selection for tasks involving closely related, low-resource languages, highlighting the strengths of regional multilingual models and the pitfalls of catastrophic forgetting.

7 Limitations

Our study, while providing valuable benchmarks, has several limitations that open avenues for future research. The scope of our analysis is confined to the Kinyarwanda-Kirundi language pair using news-domain datasets, and thus the findings may not directly generalise to other Bantu languages or text domains. A significant methodological limitation is the catastrophic forgetting observed, particularly in BantuBERTa and the traditional models, as our experimental setup did not employ mitigation strategies like continual learning; investigating such techniques is a critical next step. Finally, the performance variance between the multilingual models suggests that the scale and composition of the pre-training corpus are key. A deeper analysis of these pre-training regimens is needed to fully understand the drivers of robust cross-lingual transfer.

Acknowledgments. The authors gratefully acknowledge the support of the ABSA Chair of Data Science for facilitating this research. DSFSI is supported by the UK International Development and the International Development Research Centre, Ottawa, Canada as part of the AI for Development: Responsible AI, Empowering People Program (AI4D). DSFSI is thankful for gifts from NVIDIA, Google.org, OpenAI and Meta which enable our research.

References

1. Abdul-Mageed, M., Elmadany, A., Nagoudi, E.M.B.: ARBERT & MARBERT: Deep bidirectional transformers for Arabic. arXiv preprint arXiv:2101.01785 (2020)
2. Boudad, N., Faizi, R., Thami, R.O.H.: Cross-multilingual, cross-lingual and monolingual transfer learning for Arabic dialect sentiment classification (2023)
3. Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., Stoyanov, V.: Unsupervised cross-lingual representation learning at scale. In: Jurafsky, D., Chai, J., Schluter, N., Tetreault, J. (eds.) Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 8440–8451. Association for Computational Linguistics, Online (Jul 2020). <https://doi.org/10.18653/v1/2020.acl-main.747>, <https://aclanthology.org/2020.acl-main.747/>
4. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference **1**, 4171–4186 (10 2018), <https://arxiv.org/abs/1810.04805v2>
5. Fei, H., Li, P.: Cross-lingual unsupervised sentiment classification with multi-view transfer learning. In: Proceedings of the 58th annual meeting of the association for computational linguistics (2020)
6. Gogoulou, E., Ekgren, A., Isbister, T., Sahlgren, M.: Cross-lingual transfer of monolingual models. 2022 Language Resources and Evaluation Conference, LREC 2022 pp. 948–955 (9 2021), <https://arxiv.org/abs/2109.07348v2>
7. Karakanta, A., Dehdari, J., van Genabith, J.: Neural machine translation for low-resource languages without parallel corpora. Machine Translation **32**, 167–189 (6 2018). <https://doi.org/10.1007/S10590-017-9203-5/FIGURES/7>, <https://link.springer.com/article/10.1007/s10590-017-9203-5>
8. Kim, Y.: Convolutional neural networks for sentence classification (2014), <https://arxiv.org/abs/1408.5882>

9. Lakew, S.M., Negri, M., Turchi, M.: Low resource neural machine translation: A benchmark for five African languages (3 2020), <https://arxiv.org/abs/2003.14402v1>
10. Magueresse, A., Carles, V., Heetderks, E.: Low-resource languages: A review of past work and future challenges (6 2020), <https://arxiv.org/abs/2006.07264v1>
11. Mesham, S., Hayward, L., Shapiro, J., Buys, J.: Low-resource language modelling of South African languages (4 2021), <https://arxiv.org/abs/2104.00772v1>
12. Muhammad, S.H., Abdulmumin, I., Ayele, A.A., Ousidhoum, N., Adelani, D.I., Yimam, S.M., Ahmad, I.S., Beloucif, M., Mohammad, S.M., Ruder, S., Hourrane, O., Brazdil, P., Jorge, A., Ali, F.D.M.A., David, D., Osei, S., Bello, B.S., Ibrahim, F., Gwadabe, T., Rutunda, S., Belay, T., Messelle, W.B., Balcha, H.B., Chala, S.A., Gebremichael, H.T., Opoku, B., Arthur, S.: AfriSenti: A twitter sentiment analysis benchmark for African Languages. EMNLP 2023 - 2023 Conference on Empirical Methods in Natural Language Processing, Proceedings pp. 13968–13981 (2 2023). <https://doi.org/10.18653/v1/2023.emnlp-main.862>, <https://arxiv.org/abs/2302.08956v5>
13. Niyongabo, R.A., Hong, Q., Kreutzer, J., Huang, L.: KINNEWS and KIRNEWS: Benchmarking cross-lingual text classification for Kinyarwanda and Kirundi. In: Scott, D., Bel, N., Zong, C. (eds.) Proceedings of the 28th International Conference on Computational Linguistics. pp. 5507–5521. International Committee on Computational Linguistics, Barcelona, Spain (Online) (Dec 2020). <https://doi.org/10.18653/v1/2020.coling-main.480>, <https://aclanthology.org/2020.coling-main.480/>
14. Ogueji, K., Zhu, Y., Lin, J., Cheriton, D.R.: Small data? no problem! exploring the viability of pretrained multilingual language models for low-resource languages pp. 116–126 (2021), <https://github.com/google-research/bert/>
15. Oladipo, A., Ogundepo, O., Ogueji, K., Lin, J., Cheriton, D.R.: An exploration of vocabulary size and transfer effects in multilingual language models for African languages, https://en.wikipedia.org/wiki/Ge'ez_script
16. Parvess, J.: BantuBERTa: Using language family grouping in multilingual language modeling for bantu languages (2023), <https://repository.up.ac.za/handle/2263/92766>, available at <https://repository.up.ac.za/handle/2263/92766>
17. Parvess, J., Marivate, V., Akinyi, V.: BantuBERTa model (2024). <https://doi.org/10.57967/hf/3067>, available at <https://huggingface.co/dsfsi/BantuBERTa>
18. Pham, T., Le, K.M., Tuan, L.A., Chi, H.: UniBridge: A unified approach to cross-lingual transfer learning for low-resource languages (6 2024), <https://arxiv.org/abs/2406.09717v3>
19. Ragni, A., Knill, K., Rath, S., Gales, M.: Data augmentation for low resource languages (9 2014), https://www.isca-speech.org/archive/interspeech_2014/i14_0810.html
20. Rasooli, M.S., Farra, N., Radeva, A., Yu, T., McKeown, K.: Cross-lingual sentiment transfer with limited resources. Machine Translation **32**(1–2), 143–165 (Jun 2018). <https://doi.org/10.1007/s10590-017-9202-6>, <https://doi.org/10.1007/s10590-017-9202-6>
21. Savant, R., Shelke, A., Todmal, S., Kanphade, S., Joshi, A., Josh, R.: Universal cross-lingual text classification. In: 2024 IEEE 9th International Conference for Convergence in Technology (I2CT). pp. 1–6 (2024). <https://doi.org/10.1109/I2CT61223.2024.10543381>

Injecting Explicit Cross-lingual Embeddings into Pre-trained Multilingual Models for Code-Switching Detection

Thapelo Sindane¹, Vukosi Marivate^{1,2}, and Avashlin Moodley

¹ Data Science For Social Impact Lab, University of Pretoria, Pretoria,
sindane.thapelo@tuks.co.za,

² Lelapa AI,
vukosi.marivate@up.ac.za
³ The Council of Scientific and Industrial Research,
amoodley1@csir.co.za

Abstract. Code-switching has become the modus operandi of internet communication in many communities, such as South Africans, who are domestically multilingual. This phenomenon has made processing textual data increasingly complex due to non-standard ways of writing, spontaneous word replacements, and other challenges. Pre-trained multilingual models have shown elevated text processing capabilities in various similar downstream tasks such as language identification, dialect detection, and language family discrimination. In this study, we extensively investigate the use of pre-trained multilingual models - AfroXLMR, and Serengeti for code-switching detection on five South African languages: Sesotho, Setswana, IsiZulu, IsiXhosa, and English, with English used interchangeably with the other four languages, including various transfer learning settings. Additionally, we explore the modeling of known switching pairs within a dataset through explicit cross-lingual embeddings extracted using projection models: VecMap, Muse, and Canonical Correlation Analyses (CCA). The resulting cross-lingual embeddings are used to replace the embedding layer of a pre-trained multilingual model without additional training. Concretely, our results show that performance gains can be realized (from 59.1% monolingual to 74.1% cross-lingual, and to 90.8% multi-lingual) by closing the representational gap between the languages of the code-switched dataset with known codes, using cross-lingual representations. Moreover, expanding code-switched datasets with datasets of closely related languages improves code-switching classification, especially in cases with minimal training examples.

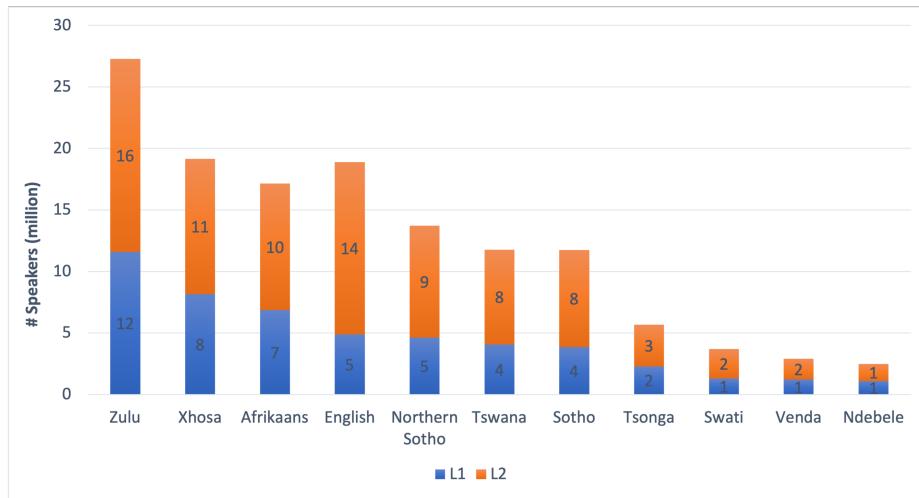
Keywords: Pre-trained Multilingual Models, Cross-lingual Embeddings, Code-switching Detection

1 Introduction

Code-switching, a term used interchangeably with code-mixing in Natural Language Processing (NLP), refers to a linguistic phenomenon where a single utterance or text is made up of multilingual tokens or words (referred to as codes), arranged meaningfully to

the receiving audience, prevalent in multilingual communities such as the Internet [18]. In a typical South African household setting, the majority of families are multilingual, as seen in Figure 1⁴. This figure highlights the number of people using each language (e.g. Zulu) as the first or preferred language, and alternatively, how many people use it as a second additional language. However, language technologies developed for these South African low-resourced languages are monolingual in nature. This myopic view limits the ability of these users to fully express themselves within these technologies, thus motivating a need for modeling code-switching in speech or text understanding technologies.

Fig. 1: Number of First Additional (L1), and Second additional Speakers (L2) in a South African Household.



Code-switching is not a new challenge. In 2013, [24] investigated the implications of code-switching for developing automated speech recognition systems using Sepedi and English code-switched datasets sourced from radio broadcasts. In [37], the authors investigated code-switching detection formalized as language detection together with Part-of-speech (POS) tagging. While [33] explored the creation of POS datasets from code-switched tweets that closely resemble real-world scenarios. On the other hand, [3] proposed a shared task for modeling Named Entity Recognition (NER) on code-switched datasets. Closer to our methodology, [17] proposed a hybrid model composed of an attention mechanism and a recurrent neural network for code-switching detection. However, none of the aforementioned works attempted a natural standpoint that brings the involved languages in the code-switched dataset closer together through cross-lingual representation learning. Cross-lingual representation learning emerges from a

⁴ https://en.wikipedia.org/wiki/Languages_of_South_Africa

broader field of cross-lingual models, where the idea is to fuse two or more syntactic and semantically sound monolingual embeddings in order to supplement the linguistic shortcomings of both embedding spaces for improving downstream task performance [23, 15, 5]. We hypothesize that given known languages $L_g = \{l_1, l_2, \dots, l_n\}$ in the labeled code-switched dataset C_{cs} , developing a language model M_L through cross-lingual representation learning of languages in L_g can improve on the task of code-switching classification (and possibly generalize to other extensions of code-switching tasks such as code-switched POS, code-switched NER, and more complex tasks like code-switched question answering) following the linguistic equivalence constraint of code-switching. Using cross-lingual representation learning to bring the languages in the code-switched datasets closer together may establish a representational space that closely matches code-switching, thus improving training, contrary to treating the codes as separate entities during training. Subsequently, we then want to answer the question: what effect do explicit cross-lingual representations generated from the involved language have on the performance of code-switching classification? To answer this, we conduct a systematic analysis and evaluation of supervised models using both monolingual and explicit cross-lingual embeddings on the task of word-level code-switching classification. The field of cross-lingual models has witnessed great success over the years in NLP and has evolved to advanced processing capabilities using large pre-trained multilingual models, where cross-lingual representation is implicitly learned [12, 10]. On this, our methodology investigates the downstream performance of code-switching classification prior to the injection of cross-lingual embeddings into the embedding layer of transformer architectures, AfroXLM-r, and Serengeti as our training and classification models. To the best of our knowledge, this work is the first study to investigate the injection of cross-lingual embeddings into pretrained multi-lingual models for code-switching detection or classification, regardless of the outcomes.

In summary, our major contributions are succinctly organized as follows:

1. We investigate the use of cross-lingual embeddings in the context of code-switching classification for four South African low-resource languages, namely, Setswana, Sesotho, isiZulu, and IsiXhosa, and one high-resourced language - English.
2. We showcase the experimental results that highlight an intuitive relationship between code-switching and cross-lingual representations, through linear improvements of code-switching detection from monolingual embeddings, to explicit cross-lingual embeddings, all the way to implicit multilingual representational training.
3. We highlight characteristics that are deterministic of improved performance for code-switching classification using cross-lingual and monolingual embeddings.

The rest of the paper is organized as follows: Section 2 discusses the related works, Section 3 outlines the methodology, broken down into the discussion of the monolingual datasets for training embeddings, an outline of our multilingual lexicons, a description of our code-switching datasets, details on pre-processing steps, techniques for generating monolingual embeddings, techniques for generating cross-lingual embeddings, details of pre-trained models used, the approach for injecting embeddings into pre-trained models, and the experimental setup for the aforementioned sub-section. Section 4 presents the experimental findings and results of this study, while Section 5 provides analyses of the findings. Finally, Section 6 gives the concluding remarks.

2 Related Work

Recent studies show the need to accelerate code-switching inclusion in language technology developments and the limitations incurred if otherwise [34]. In line with this shared responsibility, various works addressing different challenges across different downstream tasks have been proposed for code-switching. For code-switching detection, an early work by [9] proposed a Conditional Random Field (CRF) model based on their success in sequence labeling and a sub-word segmental approach that segments lexical units according to their morphological grounding. [40] explores sub-word vectors from [8], prefix and suffix extraction, and linear-chain CRF for code-switching detection and reports (at the time) promising results (83.0 %, 94.9 % F1 score, and accuracy respectively). While [27] remedies incorrect annotations of borrowed words in the code-switched text by proposing a set of likelihood metrics that utilize language usage patterns on Twitter.

Code-switching classification as a stand-alone downstream task has use-case limitations, and as such, various works have coupled it with other downstream tasks. [35] exploits existing probabilistic tree-based taggers to generate training features of the switched languages for POS tagging in the context of code-switching. They use heuristic-driven methods to combine these information-rich features to assist machine-learning-based prediction. [7] uses a joint Factorial CRF model to process complex trilingual code-mixed data for code-switching detection and POS tagging simultaneously. In contrast, [16] proposes code-switching-tailored lexical normalizers to improve the impact of non-canonical data points on POS tagging.

For code-switching in the context of NER, [6] takes an architecture-driven approach and proposes a hybrid model consisting of enriched pre-trained embeddings, a Bidirectional Long and Short Term Memory (Bi-LSTM) model to process the left and right context over the continuous representations, a convolution layer to model spatial dependencies, and finally a CRF for sequence prediction. Closer to our work, [39] concatenates English and Spanish monolingual pre-trained embeddings together with character-level representation to address the out-of-vocabulary (OOV) issue. In contrast to our work, they did not explore cross-lingual representations. In the advent of the attention era, [38] addresses both NER and code-switching detection using multilingual meta-embeddings extracted using a fully connected neural network and an attention mechanism layer.

In the aforementioned works, South African low-resourced languages lag behind across multiple tasks and techniques for processing code-switching text. In an attempt to remedy this limitation as well as ignite research interest, we propose a systematic analysis of using both monolingual and cross-lingual vector representations for code-switching detection, together with injecting static cross-lingual embeddings into the embeddings layer of the transformer architectures. In this study, we were able to show a progressive improvement over monolingual embeddings training to explicit cross-lingual embedding training, and thus implicit multilingual (i.e. of higher cross-lingual quality) embedding training, highlighting an intuitive relationship between languages involved in shared representation space, and the modeling of code-switching.

3 Methodology

3.1 Corpora

This study explores four low-resourced South African languages (LRSAL), namely, Sesotho, Setswana, isiZulu, and IsiXhosa, and one high-resourced language - English with codes: sot, tsn, zul, xho, and eng, respectively. These languages are chosen based on the availability of code-switching datasets. The sources for our monolingual corpora are: Flores [11], Conference on Machine Translation (WMT) [11], multilingual colossal clean crawled corpus (MC4) [30], National Center for Language Human Technology (NCHLT) [14], and African Crawl Dataset [36]. The statistics of our monolingual data are outlined in Table 1. This table contains the merged monolingual datasets for the four languages, where the columns 'sentences', and 'Unique vocabulary' indicate the number of sentences and unique words after the merge. The columns 'After LID', and 'Unique vocabulary' outline the number of sentences and unique vocabulary that remained after performing language identification using GlotLID [19] as mentioned under pre-processing below.

Table 1: This table outlines the monolingual datasets collected sources: Flores [11], WMT [11], MC4 [30], and NCHLT [14].

Lang.	Sentences	Unique Voc	After LID	Unique Voc
tsn	1.1M	388K	462K	118K
sot	2.6M	1.5M	750K	453K
xho	2.7M	2.8M	1.2M	1.2M
zul	7.6M	9.2M	1.5M	1.5M

3.2 Bilingual Lexicons

We used bilingual lexicons to generate the cross-lingual representation from the following sources: Cape Peninsula University of Technology (CPUT)⁵, Open Education Resource Term Bank (OERTB) [29], and our manually collected lexicons from government public school repositories⁶. The collection resulted in 8742 en-tsn, 8763 en-sot, 11117 en-xho, and 17406 en-zul bilingual lexicon pairs.

3.3 Code-Switched dataset

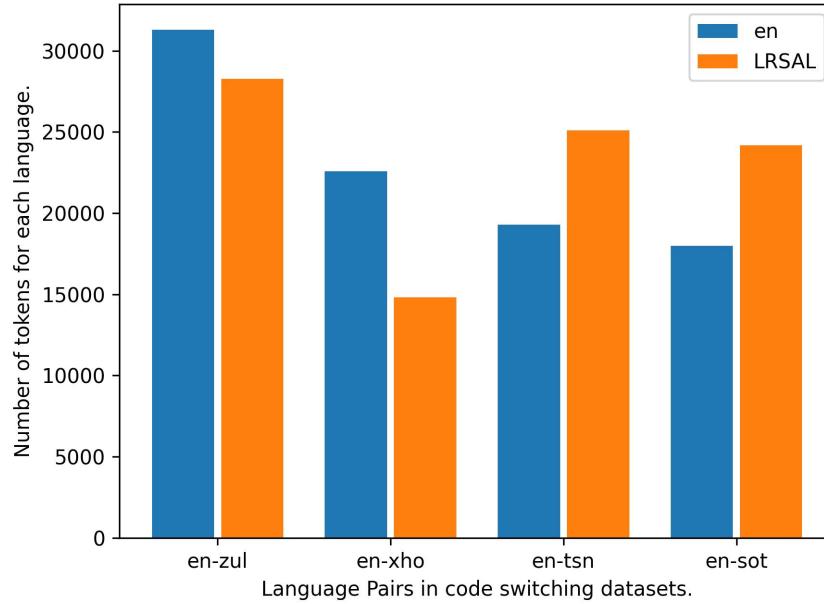
We used the labeled code-switching dataset sourced from South African soap operas [25]. This dataset covers a continuum of spontaneous code-switching types, such as intersentential, and intrasentential, and insertion. The languages covered in this dataset are as follows: English-isiZulu, English-isiXhosa, English-Sesotho, and English-Setswana.

⁵ <https://mlg.cput.ac.za/>

⁶ <https://github.com/dsfsi/za-mafoko?tab=readme-ov-file>

Figure 2 shows the switching distribution of each pair in the datasets. Concretely, the figure describes the number of tokens/words for each language in the code-switched dataset.

Fig. 2: Code distributions in code-switched datasets [25].



3.4 Pre-processing

A manual inspection of individual sources of monolingual corpora indicated that sources such as WMT, and MC4 contain foreign sentences and warrant further cleaning. For this, random samples were continuously drawn from the datasets for manual inspection. These samples were manually inspected by team members who are L1 (first language) speakers of the language. The L1 inspector is then tasked to highlight if additional filtering should be done. In most cases, the inspector would recommend filtering, and since the monolingual corpora are large, we assumed the issue persists in the datasets. As such, all datasets underwent language identification filtering. It is important to note that this process does not affect the quality of the dataset, only the size of the dataset. To address this, we used a publicly available language identification tool - GlotLID [19], which has been shown to have superior language identification performance in [32]. Since the

monolingual corpus is arranged by language, we used GlotLID to confirm if the sentence s_i of a known corpus C_{lang} belongs to language $lang$. If GlotLID identifies s_i as not of $lang$ of corpus C_{lang} , it is discarded. Table 1, shows the number of sentences from a collection of all monolingual corpora for each language, and the remaining number of lines before and after utilizing GlotLID on the corpora. Further pre-processing included the removal of URLs, numbers, punctuations, and then lower casing all words. For the annotated code-switching dataset, no further processing was done as this dataset is cleaned and word-annotated.

3.5 Monolingual Embeddings

Monolingual embeddings are continuous vector representations of words [22]. Various studies for generating these embeddings have been proposed, with the recent FastText technique [8] showing improvement over previous methods such as Word2Vec⁷, and GloVe [28] for low-resourced settings. As such we generated monolingual embeddings using FastText with dimension $d = \{50, 100, 150, 200\}$. For English, we used the available largely pre-trained embeddings from GloVe, due to having the desired 3 dimensions $d = \{50, 100, 200\}$. To evaluate the intrinsic quality of all our embeddings, we used the SimLex-999 dataset released by [21], consisting of Setswana and Sepedi paired words together with their English translations. From this, we derived that monolingual embeddings for low-resourced languages require more data to train, as reflected by their inability to capture similar words with similar representations as compared to English representations (Appendix B.1). However, projecting the inefficiently learned representations into a shared space with an effectively learned English representation space using the 3 techniques (Canonical Correlation Analyses (CCA) [15], VecMap [5], and MUSE [20]) shows improved word similarity measures between intra-similarity (within a language) and inter-similarity (between languages) (Appendix B.2). We further calculated the Spearman’s correlation of the datasets and this shows that cross-lingual embeddings outperform monolingual embeddings. An extension of this analysis is available in Appendix B.3, showing plots for all embedding dimensions and the 3 projection types to evaluate the intrinsic quality of the embeddings. We leave details in the Appendix sections as these are not the priority of this study.

Notably, lower embedding dimensions of unrelated languages (xho and zul) show higher cosine scores for paired word similarity evaluation.

3.6 Cross-lingual Embeddings

Cross-lingual embeddings are shared vector representations generated using projection techniques that aim to join monolingual embeddings of two or more languages together with the objective of transferring common desirable linguistic properties [23]. Projection techniques develop a mathematical model driven by available supervision resources, such as bilingual lexicons, parallel sentences, objective functions, etc. that aims to learn how to translate source embeddings (typically of high-resourced language) to target embeddings (of low-resourced) by transforming (shift, distort, etc.) the source embeddings into a

⁷ <https://code.google.com/p/word2vec/>

shared vector space. A comprehensive survey detailing various challenges, opportunities, future works, and applications of cross-lingual embeddings on downstream tasks can be found in [31]. In this study, we compared three pioneering projection techniques – Canonical Correlation Analyses (CCA), VecMap, and MUSE to generate cross-lingual shared representations.

3.7 Pre-trained Models

Pre-trained multilingual models such as mBERT [12], RemBERT, XLM-r [10], and their Afro-centric counterparts: Afri-BERTa [26], Afro-XLM-r [4] have shown astonishing results for many downstream tasks such as NER, POS Tagging, Machine Translation, etc., with Afro-centric methods having a slight performance edge over massively pre-trained multilingual models with minimal to no exposure to African languages. As such, we will only concentrate on the Afro-centric model Afro-XLM-r (base and large), and the recent model Serengeti [1] due to their high performance gains on downstream tasks.

3.8 Injecting pre-trained multilingual models with explicit cross-lingual embeddings

Suppose we know the languages $L_g = \{l_1, l_2, \dots, l_n\}$ of the code-switched corpus C_{cs} . We know that the interchangeable use of linguistic tokens t_i, \dots, t_k from $\{l_1, \dots, l_n\}$ is not random but rather orchestrated meaningfully to bring the targeted languages coherently together. Therefore, we hypothesize that creating shared cross-lingual representations of the known target languages that meaningfully tie the target languages semantically together may improve code-switching processing. Concretely, this study's code-switched datasets are a switching condition between *eng* and the four languages *sot*, *tsn*, *xho*, and *zul*. Hence, for known pairs (e.g *eng* – *sot*), we could create shared representations between *eng*, and *sot* resulting into a meaningful shared vector space, such that translation pairs between the two languages are semantically connected (i.e translation pairs between the two languages having similar representations), which could reflect in better processing of the target code-switching dataset such as improved handling of missing vocabulary. To evaluate this theory, we devised a technique that explores the use of cross-lingual embeddings into transformer architectures by replacing the embeddings layer with the new shared representation of the known pairs. Monolingual embeddings (to model an instance where the languages are not brought closer), and cross-lingual embeddings (to model a case where languages are brought closer together) will be used to replace the embedding layer of the transformer architectures. Experiments of these two setups will be compared and contrasted.

Dimensionality Reduction and Expansion Our cross-lingual embeddings dimensions of $d = \{50, 100, 150, 200\}$ are significantly lower than the embeddings of AfroXLMr, and Serengeti of 700+. Therefore, we are tasked with either reducing the embeddings of the transformer when combined with cross-lingual embeddings or expanding cross-lingual embeddings to match existing transformer embeddings. In this case, we experimented with randomly initialized paddings with a normal distribution to expand the cross-lingual

embeddings to match transformer dimensions. For dimensionality reduction, Principal Component Analysis (PCA), Uniform Manifold Approximation and Projection, or other techniques can be adopted to reduce the existing transformer embeddings to match the static embeddings’ shape. Due to limited space, we leave this for future work.

3.9 Experimental Design

Corpora sizes Following the pre-processing step, we extracted 80% of the most frequent words to generate our word embeddings.

Code-Switched dataset We explored various setups to divide the dataset into train, development, and test sets. The conventional setup included having a uniform train set across all languages of 4242 sentences and a test size of 1000 sentences, with only varying development sizes, since the size of the datasets for each language was not equal. The second setup included combining the dataset of the previous step to explore transfer learning. We combined code-switched datasets containing closely related languages (*xho* and *zul*) – two Nguni languages, and (*sot* and *tsn*) – two Sotho-Tswana languages. Finally, for each group, we added, a new language coming from a different language family, and then our last setup combined all languages.

Monolingual Embeddings generation Our monolingual embeddings were trained for 50 epochs, with mostly default set-ups for FastText except for min character and maximum character considerations of 1, and 5, respectively, attained empirically. Furthermore, all words were reduced to lowercase.

Cross-lingual Embeddings generation All projection techniques CCA, VecMap, and MUSE use a supervised projection setup with all available lexicons in this study. We used the default hyperparameter setup recommended by each technique’s proposed paper since no available resources exist to evaluate the intrinsic quality of cross-lingual representations.

3.10 Pre-trained Models

Each pre-trained model was trained for 20 epochs, used a batch size of 16, a maximum sequence cut-off of 200, a learning rate of $5 \exp^{-5}$ following [2], and [13]. However, Afro-XLMr-large did not perform well for the setup, and its hyperparameters were changed to a learning rate of $2 \exp^{-5}$, a batch size of 32, and was trained for 10 epochs.

4 Results

This section discusses the results of this study’s experimental findings. We report F1 score instead of accuracy since there is a clear class imbalance between the codes within our datasets (Figure 2). F1 score is the harmonic mean of the model’s precision (Equation 1) and recall (Equation 2), defined in Equation 3 below:

$$Precision = \frac{TP}{TP + FP} \quad (1)$$

$$Recall = \frac{TP}{TP + FN} \quad (2)$$

$$F1_score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (3)$$

where TP , FP , FN denote the true positive, false positive, and false negative predictions, respectively.

4.1 Baselines

Table 2, reports the F1 score performance of our baseline models: Afro-XLM-r base (b), Afro-XLM-r large (l), and Serengeti. Our results show that the models perform on par for word-based code-switching detection. Notably, increasing the dataset size by combining datasets (e.g, combining engxho with engzul to make engxhonzul) shows to have a contradicting impact on code-switching detection. Firstly, combining datasets shows a drop in performance compared to monolingual training. Secondly, training a three pair of either 2 Nguni and one Sotho-Tswana or vice versa shows a 1-point drop when training with two Nguni (*xho*, *zul*), and one Sotho-Tswana (*tsn*), compared to training with only the two Nguni languages. We hypothesize that the addition of a language from a different language family creates an interference in the internal representation largely skewed to Nguni morphology, which negatively impacts the model’s performance. Conversely, the Sotho-Tswana and single Nguni trio illustrate a performance gain in this setup. Since this pattern occurs on all models, it may imply that the internal structures of Sotho-Tswana may be robust and less susceptible to interference compared to Nguni language learning. However, we leave this investigation for future work and continue investigating transfer learning. Lastly, Afro-xlmr-base performs on par with Serengeti, while the Afro-xlmr-large model slightly lags behind with mostly 1 percentage point below across all datasets.

Table 2: Baseline model F1-scores for code-switching detection across datasets, averaged over 5 runs.

Baselines	Code-Switching dataset with base (eng)										
	xho	zul	sot	tsn	xhonzul	sottsn	xhonzulsot	sottsnul	sottsnxho	sottsnzul	xhonzulsottsn
Afro-xlmr-b	98.0	96.8	90.8	92.5	92.4	83.0	92.3	92.2	86.7	87.4	87.0
Afro-xlmr-l	96.8	96.9	89.9	89.8	90.0	81.8	85.3	91.5	85.3	86.0	87.2
Serengeti	96.8	96.9	89.9	89.8	92.4	81.8	97.2	91.4	85.3	86.0	87.2

4.2 Baselines Transfer learning

We investigate three modes of transfer learning in this section. The first setup investigates transferring models trained with multiple combinations above onto original (non-combined) datasets. The second transfer setup investigates language family grouping, where languages belonging to the same family (e.g *xho* and *zul*) are grouped into one label *nguni*, and *sttn* for Sotho-Tswana (results reported Table 3). Surprisingly, Afro-XLMR-b performs better than Serengeti on average for family-based-grouping code-switching detection. Additionally, all 3 models show higher performance for Nguni combinations over Sotho-Tswana combination datasets. In the last setup, however, we changed the datasets to only have two labels (*eng*, and *swtc*), by replacing any other label that is not *eng* to *swtc*. This is done, to investigate if transfer could be easily modeled if the label set is reduced in a multi-code switching dataset. The results for the aforementioned last experiments are reported in Table 4. From the onset, the results show performance gains on language combinations compared to baseline results in Table 2. We hypothesize that this happens because the models may find it easy to create two representational clusters of the code-switching as opposed to creating multiple representing each language. Additionally, closely related languages may not cause incorrect predictions as in the conventional setup, as these are viewed as the same class *swtc*. This is also supported by higher scores for the original dataset scores. Regardless, this approach may be a better alternative to code-switching detection as many African languages are not effectively supported at the onset (i.e. when pre-training these large multilingual models), where, instead we model code-switching as certainty (*eng*) and uncertainty (not *eng*) binary classification with the assumption that English will easily be detected with high confidence (i.e. due to high prevalence of large datasets used for pre-training). With this setup, smaller models such as Naive Bayes, SVM, etc, can be used for language identification of the *swtc* tag for finer-grained detail extraction, thus breaking the task into two subsequent tasks. However, this claim requires additional empirical evidence and we leave it for future works. Finally, on average, Serengeti performs better compared to the two variations of Afro-XLM-r for this transfer category.

Table 3: Reports model’s F1-score for code-switching detection for each dataset averaged over 5 runs of grouped label datasets. The *xho*, *zul* are grouped into *Nguni*, and *sot tsn* grouped into *Sotho-Tswana* (*sttsn*).

Transfer	Code-Switching dataset with base (<i>eng</i>) and labels <i>nguni</i> and <i>sttsn</i>						
	<i>xhozul</i>	<i>sottsn</i>	<i>xhozulsot</i>	<i>xhozultsn</i>	<i>sottsnxho</i>	<i>sottsnzul</i>	<i>xhozulsottsn</i>
Afro-xlmr-b	97.8	92.4	94.9	95.7	93.8	93.8	94.6
Afro-xlmr-l	97.4	91.9	94.9	95.6	93.6	93.4	94.3
Serengeti	97.4	91.9	94.9	95.7	93.6	93.4	94.3

Table 4: Reports the label change transfer model’s F1-score for code-switching detection for each dataset averaged over 5 runs.

Transfer	Code-Switching dataset with base (eng) and second label switched (swtc)						
	xhozul	sottsn	xhozulsot	xhozultsn	sottsnxho	sottsnzul	xhozulsottsn
Afro-xlmr-b	85.9	93.1	95.4	96.0	93.6	94.0	95.1
Afro-xlmr-l	97.9	92.5	95.5	96.1	94.1	80.7	95.1
Serengeti	97.6	92.2	95.1	95.9	93.8	93.7	94.8

4.3 Injecting pre-trained multilingual models with explicit cross-lingual embeddings

Table 5, shows the experimental results of this study using various cross-lingual embeddings settings. Due to limited space, we only reported the results of Afro-XLM-r-base. We presented the remaining results in Appendix C with accompanying discussions and the Serengeti model results. From these results, CCA and VecMap embeddings show better performance for lower dimensions 50 and 100 over Muse embeddings, while Muse surpasses these two techniques on the highest dimension 200 on the original datasets *xho*, *zul*, *sot*, and *tsn*. This behavior is not clear as to why it occurs, as we expected Muse embeddings to perform consistently better than the other two techniques, as shown by its representation quality illustrated in the intrinsic evaluation (Appendix B.2, Appendix B.3). This could mean that intrinsic performance is not correlated to extrinsic performance for this task. On the combination datasets, only Muse embeddings show consistent performance, while the performance of CCA and VecMap fluctuates depending on the data combinations. This could be due to that, while the English to low-resourced language ($en - LRL_i$) pair remains semantically connected through the projection, the inclusion of a foreign pair $en - LRL_k$ in the same model introduces the same issue (i.e. the disconnect) we are trying to solve in the embedding space (i.e. the attempt to bring representation of the code-switched languages closer together), for i , and k being low-resource languages. For example, combining *en-tsn*, with *en-xho* datasets, results in interference between *tsn* and *xho* as these were not explicitly joined through cross-lingual projection techniques. Therefore, there is no common space for the two pairs, thus creating the very issue we are trying to solve.

To understand the impact of these injected cross-lingual embeddings we trained the best-performing model - Afro-XLMr-base, with monolingual embeddings from which these joined representations were formed. Table 16, shows these results for non-projected embeddings of *xho*, *zul*, *sot*, *tsn* created using FastText. From this, we can see that monolingual performance is significantly lower for all of the various evaluation metrics: Accuracy, Precision, Recall, and F1-score. This, in a way, justifies our hypothesis that, perhaps not treating the codes as independent spaces but rather semantically joining their individual spaces at pre-training could improve code-switching processing as these languages are brought to the same shared space.

From these results collectively, it is plausible that the high performance of baseline pre-trained models can be due to the implicit multi-lingual representation learned over vast amounts of textual data compared to the small samples used to train our cross-lingual

Table 5: Reports the Afro-XLMr-base F1-score for code-switching detection for each dataset, averaged over 5 runs using the three embedding techniques. The injected embeddings were randomly selected from the full embedding matrix to match the vocabulary size of the transformer.

Models	Code-Switching dataset with base (eng)										
	xho	zul	sot	tsn	xhozul	sottsn	xhozulsot	xhozultsn	sottsnxho	sottsnzul	xhozulsottsn
Embedding Dimension: 50											
CCA	81.5	84.6	77.1	79.5	57.1	56.7	73.8	61.5	66.5	77.6	78.7
VecMap	75.7	84.0	85.3	83.9	80.7	64.7	83.2	76.4	67.1	63.1	45.0
Muse	76.2	79.6	62.8	68.1	70.5	70.4	73.1	74.3	59.3	62.1	69.7
Embedding Dimension: 100											
CCA	81.3	83.2	76.0	66.5	69.1	74.8	74.1	75.8	79.6	78.2	56.5
VecMap	81.5	84.1	78.1	84.7	69.4	74.3	82.5	45.7	75.5	68.0	62.0
Muse	67.0	81.0	76.1	74.1	72.1	70.9	74.2	64.4	70.4	71.9	70.0
Embedding Dimension: 200											
CCA	71.3	61.2	77.6	71.8	68.1	66.8	48.5	71.3	79.7	72.1	59.7
VecMap	71.2	84.0	79.5	75.7	72.5	75.2	79.3	69.6	78.1	59.4	77.9
Muse	76.1	79.6	76.6	74.9	71.4	71.1	74.4	75.5	71.9	71.7	71.7

Table 6: Reports the Afro-XLMr-base Accuracy (acc), Precision (prec), Recall (rec), and F1-score (f1) for code-switching detection for each dataset averaged over 5 runs using FastText Monolingual embeddings and Glove embeddings for English words. The injected embeddings were randomly selected from the full embedding matrix to match the vocabulary size of the transformer.

Metric	Code-Switching dataset with base (eng)										
	xho	zul	sot	tsn	xhozul	sottsn	xhozulsot	xhozultsn	sottsnxho	sottsnzul	xhozulsottsn
Embedding Dimension: 50											
acc	93.4	81.0	90.8	90.8	80.5	82.4	86.8	82.8	66.7	76.2	78.6
prec	71.4	56.3	60.5	68.8	55.6	63.5	63.4	60.6	47.4	57.1	59.3
rec	68.1	52.5	61.7	64.9	56.2	62.4	64.2	61.3	40.5	54.3	61.3
f1	69.1	53.6	61.1	66.6	55.7	62.3	63.6	60.7	42.7	55.1	59.9
Embedding Dimension: 100											
acc	92.9	86.7	71.8	78.6	81.5	75.5	81.3	74.0	78.4	76.2	75.0
prec	63.2	57.7	57.4	60.5	58.3	54.8	55.4	52.4	58.9	58.8	54.6
rec	63.6	55.9	44.3	54.1	59.7	55.4	49.4	51.2	61.3	54.7	51.9
f1	63.4	56.3	49.1	56.9	58.7	54.6	51.6	51.3	59.7	55.8	52.4
Embedding Dimension: 200											
acc	92.4	80.9	75.7	90.6	72.2	82.4	65.3	65.0	55.3	81.5	75.3
prec	66.4	56.2	56.0	71.8	46.1	62.6	45.1	43.7	36.5	61.5	56.3
rec	64.1	52.4	32.5	66.2	44.1	61.6	40.6	39.7	27.7	62.8	53.5
f1	64.6	53.5	41.1	68.7	44.7	61.4	42.1	40.9	30.7	61.7	54.1

embeddings. Regardless, a clear trend emerges, that from monolingual embeddings, a boost in performance can be attained by using cross-lingual embeddings, and further improvements is attained by using largely trained multilingual embeddings.

5 Analyses and Discussion

We also aimed to investigate the extent to which shared representations support text processing, especially for unknown or out-of-vocabulary words for improved performance. That is, in cases where vocabulary is missing, we want to analyze if other related words are used for improving performance and if this is traced to shared representations. We aimed to achieve this through embedding attention score analyses. This is done by plotting embedding attention scores of predicted sentences to see where priority is placed for certain words.

Sentence attention score analyses consider relationships of words within a sentence, we also wanted to consider alternative words within the entire embedding matrix that may be important in processing the final output of the layer. We observed that high attention was also given to words outside the main sentence vocabulary. This could mean that, related words, were identified (possibly made possible by shared representations), thus, improving the processing of the target text. Monolingual plots do not show this behavior of exterior attention. This implies that, indeed, deeper connections may have been forged through semantically connecting monolingual embeddings. Further analyses on this is provided in Appendix D.

6 Conclusion

Code-switching has gained research attention in the field of Natural Language Processing, especially for low-resourced languages due to most communities being largely multilingual. Large pre-trained multilingual models are typically a de facto standard processing tool for many downstream tasks due to their increased processing capabilities. However, the use of explicit cross-lingual embeddings to bridge the gap between the language representations of the known codes lags behind. In this study, we explore the use of cross-lingual embeddings that aim to bring known language pairs closer together to efficiently learn code-switching detection. Indeed, our experimental analyses show that mapping the switched languages into a single shared vector space before training shows untapped processing capabilities for code-switching detection. Concretely, fine-tuning AfroXLM-r, and Serengeti architectures with explicit cross-lingual embeddings outperforms monolingual (only joined through concatenation) embeddings. Although the cross-lingual injection experiments performed poorly compared to the original multilingual embedding baselines, our results simply imply that, learning explicit shared representation between known codes, formed a high-order representational space, enhancing inter-learning between token subspaces, allowing efficient processing of code-switched text.

Limitations

The vocabulary of the embeddings is larger than the transformer embeddings. This means the selection of the appropriate words to add to the model embedding becomes crucial.

We explored a theoretically ideal scenario for generating cross-lingual embeddings and have not explored set-ups such as how many bilingual lexicons' signals are sufficient for generating the best shared representations.

In line with the above limitation, this study did not explore hyperparameter fine-tuning for injected cross-lingual embeddings. We advise future works to consider this as it may significantly improve outcomes.

Our experiments did not consider the many massively pre-trained multilingual models, such as RemBert, mBERT, including afro-centric pre-trained models such as AfroLM, AfriBerta, etc. This is due to limitations in computational power as well as time constraints as recreating and injecting explicit cross-lingual embeddings takes more time, even in the advent of multiprocessing capabilities to speed up the process.

Ethics Statement

Data and Models Disclaimer The datasets and models used in this study are collected from publicly available resources with no potential harm, threats, and risks to society.

License

This document and all its artifacts is licensed under the NOODL Equitable Data License. To view a copy of this license, visit: <https://licensingafricandatasets.com/nwulite-obodo-license>

Acknowledgements

The authors gratefully acknowledge the support of the ABSA Chair of Data Science and the Data Science for Social Impact (DSFSI) Lab at the University of Pretoria. This work was supported by UK International Development and the International Development Research Centre (IDRC), Ottawa, Canada, under the AI4D Africa Program. DSFSI also acknowledges gifts from NVIDIA, Google.org, OpenAI and Meta.

References

1. Adebara, I., Elmadany, A., Abdul-Mageed, M., Alcoba Inciarte, A.: SERENGETI: Massively multilingual language models for Africa. In: Findings of the Association for Computational Linguistics: ACL 2023. pp. 1498–1537. Association for Computational Linguistics, Toronto, Canada (Jul 2023). <https://doi.org/10.18653/v1/2023.findings-acl.97>, <https://aclanthology.org/2023.findings-acl.97>

2. Adelani, D.I., Abbott, J., Neubig, G., D'souza, D., Kreutzer, J., Lignos, C., Palen-Michel, C., Buzaaba, H., Rijhwani, S., Ruder, S., et al.: Masakhaner: Named entity recognition for african languages. *Transactions of the Association for Computational Linguistics* **9**, 1116–1131 (2021)
3. Aguilera, G., AlGhamdi, F., Soto, V., Diab, M., Hirschberg, J., Solorio, T.: Named entity recognition on code-switched data: Overview of the calcs 2018 shared task. arXiv preprint arXiv:1906.04138 (2019)
4. Alabi, J.O., Adelani, D.I., Mosbach, M., Klakow, D.: Adapting pre-trained language models to African languages via multilingual adaptive fine-tuning. In: Proceedings of the 29th International Conference on Computational Linguistics. pp. 4336–4349. International Committee on Computational Linguistics, Gyeongju, Republic of Korea (Oct 2022), <https://aclanthology.org/2022.coling-1.382>
5. Artetxe, M., Labaka, G., Agirre, E.: A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. arXiv preprint arXiv:1805.06297 (2018)
6. Attia, M., Samih, Y., Maier, W.: Ghht at calcs 2018: Named entity recognition for dialectal arabic using neural networks. In: Proceedings of the Third Workshop on Computational Approaches to Linguistic Code-Switching. pp. 98–102 (2018)
7. Barman, U., Wagner, J., Foster, J.: Part-of-speech tagging of code-mixed social media content: Pipeline, stacking and joint modelling. In: Proceedings of the second workshop on computational approaches to code switching. pp. 30–39 (2016)
8. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. *Transactions of the association for computational linguistics* **5**, 135–146 (2017)
9. Chittaranjan, G., Vyas, Y., Bali, K., Choudhury, M.: Word-level language identification using crf: Code-switching shared task report of msr india system. In: Proceedings of the first workshop on computational approaches to code switching. pp. 73–79 (2014)
10. Conneau, A., Khaderwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., Stoyanov, V.: Unsupervised cross-lingual representation learning at scale. arXiv preprint arXiv:1911.02116 (2019)
11. Costa-jussà, M.R., Cross, J., Çelebi, O., Elbayad, M., Heafield, K., Heffernan, K., Kalbassi, E., Lam, J., Licht, D., Maillard, J., et al.: No language left behind: Scaling human-centered machine translation. arXiv preprint arXiv:2207.04672 (2022)
12. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
13. Dione, C.M.B., Adelani, D., Nabende, P., Alabi, J., Sindane, T., Buzaaba, H., Muhammad, S.H., Emezue, C.C., Ogayo, P., Aremu, A., et al.: Masakhapos: Part-of-speech tagging for typologically diverse african languages. arXiv preprint arXiv:2305.13989 (2023)
14. Eiselen, R., Puttkammer, M.J.: Developing text resources for ten south african languages. In: LREC. pp. 3698–3703. Citeseer (2014)
15. Faruqui, M., Dyer, C.: Improving vector space word representations using multilingual correlation. In: Proceedings of EACL (2014)
16. Van der Goot, R., Çetinoğlu, Ö.: Lexical normalization for code-switched data and its effect on pos-tagging. arXiv preprint arXiv:2006.01175 (2020)
17. Hussain, A., Arshad, M.U.: An attention based neural network for code switching detection: English & roman urdu. arXiv preprint arXiv:2103.02252 (2021)
18. Jose, N., Chakravarthi, B.R., Suryawanshi, S., Sherly, E., McCrae, J.P.: A survey of current datasets for code-switching research. In: 2020 6th international conference on advanced computing and communication systems (ICACCS). pp. 136–141. IEEE (2020)
19. Kargaran, A.H., Imani, A., Yvon, F., Schütze, H.: Glotlid: Language identification for low-resource languages. In: The 2023 Conference on Empirical Methods in Natural Language Processing (2023), <https://openreview.net/forum?id=dl4e3EBz5j>

20. Lample, G., Conneau, A., Denoyer, L., Ranzato, M.: Unsupervised machine translation using monolingual corpora only. arXiv preprint arXiv:1711.00043 (2017)
21. Makgatho, M., Marivate, V., Sefara, T., Wagner, V.: Training cross-lingual embeddings for setswana and sepedi. arXiv preprint arXiv:2111.06230 (2021)
22. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)
23. Mikolov, T., Le, Q.V., Sutskever, I.: Exploiting similarities among languages for machine translation. arXiv preprint arXiv:1309.4168 (2013)
24. Modipa, T.I., De Wet, F., Davel, M.H.: Implications of sepedi/english code switching for asr systems (2013)
25. Niesler, T., et al.: A first south african corpus of multilingual code-switched soap opera speech. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018) (2018)
26. Ogueji, K., Zhu, Y., Lin, J.: Small data? no problem! exploring the viability of pretrained multilingual language models for low-resourced languages. In: Proceedings of the 1st Workshop on Multilingual Representation Learning. pp. 116–126. Association for Computational Linguistics, Punta Cana, Dominican Republic (Nov 2021), <https://aclanthology.org/2021.mrl-1.11>
27. Patro, J., Samanta, B., Singh, S., Basu, A., Mukherjee, P., Choudhury, M., Mukherjee, A.: All that is english may be hindi: Enhancing language identification through automatic ranking of likeliness of word borrowing in social media. arXiv preprint arXiv:1707.08446 (2017)
28. Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). pp. 1532–1543 (2014)
29. of Pretoria, U.: Open educational resource term bank, https://www.up.ac.za/african-languages/news/post_2728581-open-educational-resource-term-bank-pg2
30. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. arXiv e-prints (2019)
31. Sindane, T., Marivate, V., Modupe, A.: Cross-lingual embedding methods and applications: A systematic review for low-resourced scenarios. Natural Language Processing Journal **12**, 100157 (2025). <https://doi.org/https://doi.org/10.1016/j.nlp.2025.100157>, <https://www.sciencedirect.com/science/article/pii/S2949719125000330>
32. Sindane, T.A., Marivate, V.: From n-grams to pre-trained multilingual models for language identification. In: Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities. pp. 229–239 (2024)
33. Singh, K., Sen, I., Kumaraguru, P.: A twitter corpus for hindi-english code mixed pos tagging. In: Proceedings of the sixth international workshop on natural language processing for social media. pp. 12–17 (2018)
34. Sitaram, S., Chandu, K.R., Rallabandi, S.K., Black, A.W.: A survey of code-switched speech and language processing. arXiv preprint arXiv:1904.00784 (2019)
35. Solorio, T., Liu, Y.: Part-of-speech tagging for english-spanish code-switched text. In: Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing. pp. 1051–1060 (2008)
36. Vegi, P., J. S., Paul, B., Mishra, A., Banjare, P., K R, P.K., Viswanathan, C.: Webcrawl african : A multilingual parallel corpora for african languages. In: Proceedings of the Seventh Conference on Machine Translation. pp. 1076–1089. Association for Computational Linguistics, Abu Dhabi (December 2022), <https://aclanthology.org/2022.wmt-1.105>
37. Vyas, Y., Gella, S., Sharma, J., Bali, K., Choudhury, M.: Pos tagging of english-hindi code-mixed social media content. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). pp. 974–979 (2014)

38. Winata, G.I., Lin, Z., Fung, P.: Learning multilingual meta-embeddings for code-switching named entity recognition. In: Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019). pp. 181–186 (2019)
39. Winata, G.I., Wu, C.S., Madotto, A., Fung, P.: Bilingual character representation for efficiently addressing out-of-vocabulary words in code-switching named entity recognition. arXiv preprint arXiv:1805.12061 (2018)
40. Xia, M.X.: Codeswitching language identification using subword information enriched word vectors. In: Proceedings of the second workshop on computational approaches to code switching. pp. 132–136 (2016)

Part V

Volume II: Socio-technical and Human-Centred AI

Small Language Models on the Edge for Real-World Agentic Systems in Industry

Edward B. Duffy¹, David Fernandez², Alta de Waal³, and Mert D. Pesé²

¹ BMW Group, Information Technology Research Center, SC, USA

² Clemson University, SC, USA

³ BMW IT Hub, South Africa

Abstract. Large Language Models face significant deployment challenges in enterprise environments, including high computational costs, data privacy concerns, and network dependencies. This paper presents a framework for deploying Small Language Models (SLMs) with fewer than 7 billion parameters on edge devices, using agentic architectures to overcome capacity limitations. We introduce three key contributions: (1) a multi-agent benchmarking framework employing role-based evaluation to reduce bias, (2) a three-phase task planning pipeline that decomposes planning into subtask identification, dependency reasoning, and schema-constrained generation, and (3) real-world implementations achieving 3-4x latency improvements over cloud services. Our evaluation demonstrates that models like Phi-4 achieve CEFR C1-level translation quality and 0.883 G-Eval summarization scores on commodity hardware. Through WebLLM browser-based inference and local hosting, we show that SLMs effectively serve enterprise needs in privacy-sensitive, bandwidth-constrained, or air-gapped environments, representing a viable alternative prioritizing data sovereignty and cost efficiency.

Keywords: Small Language Models, Edge Computing, Agentic AI, Multi-Agent Systems, LLM Evaluation, On-Device Inference

1 Introduction

Increasing interest in agentic AI [16], systems that autonomously plan, reason and act, has redefined how Artificial Intelligence (AI) can be applied in enterprise environments. Rather than relying on monolithic end-to-end models, agentic AI frameworks decompose problems into smaller tasks, coordinate tools, and iterate to reach goals. These systems mirror human problem solving by using components such as task planners, memory retrieval modules, and tool users [33].

The deployment of such intelligent systems has been accelerated by Large Language Models (LLMs) such as GPT-4 [32], Claude [6] and Gemini [39], which offer impressive generalization and reasoning capabilities across diverse domains [34]. However, these capabilities come at a cost. LLMs require substantial computational infrastructure for both training and inference, and API-based access can result in expenses that exceed hundreds of thousands of dollars annually [14,31]. These financial barriers limit accessibility for many organizations.

Although agentic AI is typically powered by large foundation models, recent developments show that it can also be implemented using smaller, specialized components [49] making it an attractive architecture for enterprise environments.

Privacy and data governance concerns pose additional adoption challenges for enterprise AI systems. LLMs typically rely on cloud-based APIs, requiring sensitive information to be transmitted to third-party providers. This introduces potential compliance risks, particularly in regulated industries such as finance, healthcare, and manufacturing. Furthermore, jurisdictions such as the European Union (GDPR) [1] and South Africa (POPIA) [38] enforce strict data residency and sovereignty requirements, complicating cloud-based LLM usage.

Edge computing presents an alternative deployment approach that can address many of these challenges[50]. By processing data closer to its source, edge devices can reduce latency, minimize bandwidth requirements, and keep sensitive information within organizational boundaries [28]. However, traditional edge hardware has limited computational resources compared to cloud infrastructure, making it unsuitable for running large language models efficiently.

Self-hosting large models such as Llama 70B [2] or Falcon 180B [3] is not a simple alternative. These models require high-end GPU clusters and specialized infrastructure [4], making them inaccessible to smaller organizations or those operating in environments with limited connectivity and bandwidth.

In response to these challenges, Small Language Models (SLMs) have emerged as a promising alternative. These models typically contain fewer than 7 billion parameters [19] and are small enough to run on commodity hardware such as standard workstations, industrial PCs or even edge devices. This allows organizations to enable edge computing and reduces reliance on external infrastructure. SLMs open new opportunities for localized, real-time AI without compromising privacy or budget. SLMs offer several advantages for enterprise use:

- **Data Privacy:** By running locally, SLMs ensure that sensitive information remains within the organization’s network.
- **Cost Efficiency:** These models can run on existing hardware, eliminating the need for high-end GPUs and costly API usage.
- **Reduced Network Dependency:** SLMs support real-time inference in bandwidth constrained settings such as manufacturing floors.
- **Edge Deployment Capability:** SLMs can be deployed directly on edge devices, enabling distributed AI processing closer to data sources while maintaining low latency and reducing dependence on centralized infrastructure.

However, the reduced size of SLMs leads to a decrease in general reasoning, memory capacity, and instruction-following capabilities compared to state-of-the-art LLMs [13]. This limitation restricts their effectiveness for complex tasks and presents a challenge for deploying SLMs in production environments.

To address this, we introduce an approach that combines agentic principles with specialized SLMs. Instead of relying on a single large model, we implement agent frameworks that decompose tasks into smaller subtasks, each handled by dedicated SLMs. These modular agents may include planners, retrievers, or

tools users. This structure allows SLMs to operate efficiently in low-resource environments while maintaining the capability to execute complex goals.

Our methodology includes a novel benchmarking pipeline that simulates real-world workflows using agent roles such as Proctor, Student, and Grader. We evaluate SLM performance on translation and summarization tasks and analyze how specialized agent roles can enhance effectiveness despite limited model size.

Our experiments show that agentic workflows allow SLMs to achieve competitive performance in multilingual translation and summarization, with gains in privacy, and cost-efficiency. This paper makes the following contributions:

- **Novel Multi-Agent Benchmarking Framework:** We introduce an evaluation methodology that uses multiple agents in different roles (Proctor, Student, Grader) to assess SLM capabilities. This framework reduces evaluation bias through multi-perspective assessment and provides more robust performance metrics than traditional single-model evaluation approaches.
- **Three-Phase Task Planning Pipeline:** We develop a decomposed approach to task planning specifically designed for the limitations of SLMs. By breaking the planning process into subtask identification, dependency reasoning, and schema-constrained generation phases, we show significant improvements in SLM planning reliability compared to single-pass approaches.
- **Real-World Use Case Implementation:** We demonstrate the practical applicability of our approach through a multilingual translation system comparing SLM against cloud-based services, showing 3-4x latency improvements while maintaining data sovereignty and reduced network usage.

2 Related Work

2.1 Small Language Models

Recent research has demonstrated that models with fewer than 7 billion parameters can achieve competitive performance through improved training strategies [19] and high-quality data [40], while compression techniques like quantization [20] and low-rank adaptation [9] enable resource-constrained deployment. However, smaller models exhibit reduced reasoning depth and instruction-following reliability [13], particularly for complex multi-step tasks. Rather than matching LLM capabilities through better training alone, our work addresses these limitations through architectural decomposition. Our three-phase task planning pipeline (subsection 3.1) targets planning weaknesses by breaking monolithic reasoning into manageable stages: subtask identification, dependency reasoning, and schema-constrained generation.

2.2 Agentic AI Frameworks

Agentic AI frameworks like AutoGPT [45] and LangChain demonstrate multi-step reasoning and tool usage [41], with strategies such as ReAct [47] interleaving reasoning and acting, and Tree of Thoughts [46] enabling deliberate

problem-solving. Multi-agent collaboration approaches [26] leverage debate between agents for complex tasks. However, these frameworks target large models and assume sufficient individual agent reasoning capacity. We extend agentic principles to edge computing by designing orchestration strategies for resource-constrained SLMs. Our agent framework (subsection 3.1) distributes cognitive load across specialized components rather than relying on individual agent sophistication, enabling complex execution on edge-deployed smaller models.

2.3 LLM-as-a-Judge and Evaluation

The LLM-as-a-Judge paradigm [48] enables scalable model evaluation using LLMs [27], refined through systems like PandaLM [43], JudgeLM [51], and Auto-J [25]. However, single-judge systems suffer from position bias [42], length bias [36], and evaluation inconsistency [21]. Our multi-agent benchmarking framework (section 5) addresses these biases by using multiple grader agents with different LLM backends, forcing diverse reasoning paths and preventing consensus artifacts from identical model evaluations.

2.4 Edge Computing for AI

Edge AI deployment addresses latency, bandwidth, and privacy concerns by processing data near its source [50]. Existing work has focused on optimization techniques like pruning [17], quantization [20], and knowledge distillation [18] primarily for traditional deep learning models. We demonstrate practical edge deployment of language models through WebLLM [35] for browser-based inference and local hosting on commodity hardware.

2.5 Multi-Modal and Domain-Specific Applications

Recent advances have explored specialized applications of language models in various domains including legal text analysis [15], medical Q&A [37], and financial analysis [44]. Recent work in autonomous vehicles has explored Vision Language Models for crash analysis [12], adversarial traffic sign attacks [5], and LLM-based anomaly detection [11]. Our work contributes a general framework adaptable across domains while maintaining edge deployment benefits.

3 Background

3.1 Agents

An AI Agent is an autonomous software entity that perceives its environment, reasons, and acts to achieve specific goals through adaptive behavior rather than predefined scripts. As shown in Figure 1, an agent consists of several interconnected components that work together to enable intelligent behavior: a prompt

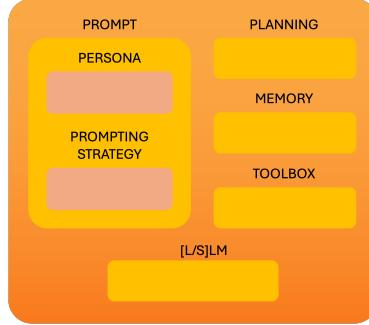


Fig. 1: **Anatomy of Agent:** Prompt layer defines the agent’s behavior through persona definition and prompting strategy. Planning component enables task decomposition and execution strategy formulation.

defining persona and strategy, autonomy and planning capabilities for multi-step decision-making, memory for context retention and learning, and tools like semantic search and APIs for enhanced functionality beyond text generation.

An Agent Framework, Figure 2, is a structured system designed to manage and coordinate AI agents to solve complex tasks. At its core is a central Coordinator, which oversees the Task Queue, a dynamic list of tasks that guides the agent’s workflow. The framework operates in three main steps: **Planning and task decomposition** (1) , where the Coordinator breaks down the user’s original query into smaller, manageable subtasks that populate the task queue. The **Executing Agents** phase (2) involves AI agents processing these tasks, leveraging tools such as APIs or external systems as needed, to generate results. Finally, the **Verification** step (3) ensures that outputs align with the intent of the original user query, validating the accuracy and completeness of the solution. This modular approach allows the framework to maintain organization, adapt to complex problems, and ensure high-quality outcomes.

3.2 SLM Task Planning

The premise of task planning is to decompose a user’s query into a set of smaller, manageable subtasks that can be distributed across specialized agents. These subtasks are typically interdependent and should not be treated as a flat, sequential list. Instead, they form a **dependency graph** where some tasks may be executed in parallel, while others depend on the completion prerequisite sub-tasks. Effective task planning requires both structured reasoning and the ability to generate outputs in a format that downstream agents or orchestrators can interpret. For this purpose, we enforce a parsable schema-defined output format, which ensures consistency and interoperability across agent components.

LLMs are capable of performing this end-to-end task with a well-tuned system prompt. A single agent can generate task breakdowns, assign dependencies, and output a schema-compliant task list. However, SLMs struggle with this approach due to limitations in context window size, reasoning depth, and memory.

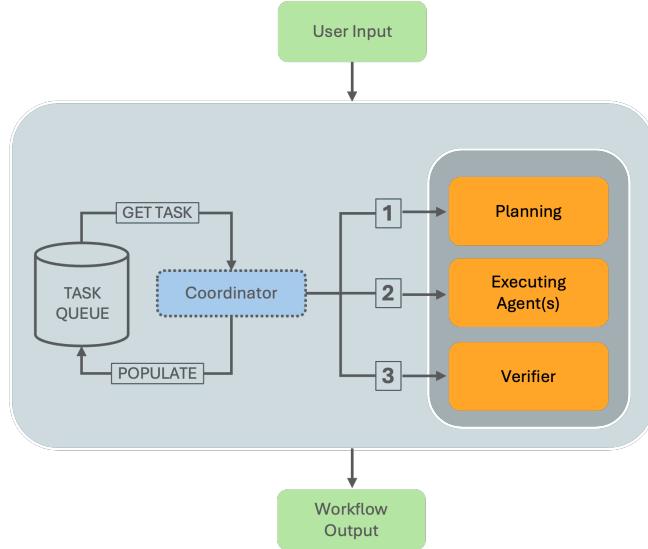


Fig. 2: **Agent Framework:** User input flows into the framework where the Coordinator manages the Task Queue and orchestrates the three-phase process.

To address these limitations, we designed a three-phase task planning pipeline tailored to SLMs capabilities. Instead of requiring the model to reason, reflect, and output data in a single call, we decompose the process into distinct stages:

1. **Subtask Identification:** The SLM is prompted to break the original query into a list of subquestions or subgoals. This stage focuses entirely on understanding and dissecting the user's intent; no output schema is enforced.
2. **Dependency Reasoning:** For each subtask, the SLM is asked to specify any other subtasks that must be completed first. This builds a dependency tree and makes the model to reason about prerequisite relationships.
3. **Schema-Constrained Task List Generation:** Finally, the SLM is prompted to generate a fully structured task list object that conforms to a predefined JSON schema. This step enforces strict formatting, improving compatibility with task orchestration engines or downstream agents.

This modular design enables SLMs to focus on a single cognitive step at a time: identifying the task, reasoning through dependencies, and formatting the final output. We found that this sequential prompting strategy substantially improves the reliability of the outputs compared to single-pass generation.

4 Methodology

Our research employs a three-phase methodology to evaluate SLMs:

Phase 1: Multi-Agent Benchmarking Framework (section 5). We developed an evaluation architecture using role-based agents, a Proctor coordi-

nates assessment, Student agents (SLMs under test) perform tasks, and multiple Grader agents with different LLM backends conduct collaborative evaluation.

Phase 2: Systematic Performance Evaluation We assessed multiple SLMs (3-7B parameters) across translation tasks (subsection 5.1) using CEFR proficiency standards and summarization tasks (subsection 5.2) using G-Eval metrics [32]. Models were evaluated under different deployment conditions (WebLLM and local hosting using V100) to measure both task quality and operational feasibility including latency, memory requirements, and offline capability.

Phase 3: Production Deployment Validation (section 6) We implemented systems to validate benchmarking results, including an Afrikaans translation service comparing edge-deployed MADLAD-400 (3B) against the cloud-based Vulavula API, measuring latency and data sovereignty maintenance.

While our evaluation primarily demonstrates translation and summarization tasks, these represent generic capabilities applicable across diverse industries (manufacturing, healthcare, finance, legal services) without requiring exposure of sensitive domain-specific data. We have also developed production systems for document comparison (legal compliance verification) and multi-step task planning, though space constraints precluded their inclusion.

5 Benchmarking

To evaluate the capabilities of various SLMs across different tasks, we designed a multi-agent benchmarking framework inspired by real-world evaluation scenarios. This framework simulates a student examination process using multiple coordinated agents, each with a clearly defined role. As shown in Figure 3, the framework operates the following components and process flow:

The **Proctor** agent, backed by an LLM, serves as the central coordinator managing the entire evaluation process. The workflow begins when the Proctor receives an **Exam** from the available tests ①. The proctor then administers this exam to the **Student** agent (SLM under test), by sending the exam prompt ②. After receiving the student’s response ③, the Proctor forwards both the original exam and the student’s translation to the first **Grader agent** ④. The first grader agent, backed by GPT, then shares the information with the second **Grader agent**, backed by Claude, to initiate a collaborative evaluation process.

The two Grader agents engage in structured discussion to evaluate student performance ⑤, ensuring objective assessment through distributed reasoning. Upon reaching consensus on grade and review, they return their evaluation to the Proctor ⑥, who records results in the Benchmark Results ⑦.

This multi-grader approach reduces bias by instantiating the Grader agents using different LLM backends, avoiding immediate agreement due to identical reasoning paths and enforcing richer evaluation of the SLM’s capabilities.

5.1 Translation Benchmarking

The translation benchmark begins by fetching the most recent English-language articles from Wikipedia’s featured article of the day section. These articles are se-

lected for their clarity, well-written structure, and topical diversity, which makes them ideal candidates for assessing multilingual translation capabilities.

The student (SLM) is instructed to produce a full translation of the article in a target language without additional commentary. The grader agents then receive both the source and translated versions and are prompted to engage in a dialog to determine the translation’s quality. The grading system follows the Common European Framework of Reference for Languages (CEFR) [10], which defines seven levels of language proficiency:

- F: Student has no understanding of the language.
- A1: Beginner level with a working vocabulary of about 700 words.
- A2: Pre-intermediate level; users understand basic expressions.
- B1: Intermediate level; users can communicate reasonably well.
- B2: Advanced level; users can communicate easily and spontaneously.
- C1: Proficient level; users can perform complex tasks.
- C2: A mastery or proficiency level; users can understand almost everything.

Table 1: Translation Benchmark Results: Average CEFR scores across languages

Deployment	Model	German	Spanish	Portuguese	Magyar	Japanese	French
WebLLM	Llama-3.2	2.900	2.940*	3.300**	1.580*	1.200	3.120
	Phi-3-mini-128k	3.040	2.340	2.860	0.340	1.200	2.098
	Qwen2.5 (f16)	3.100*	2.660	2.960	0.540	2.020*	3.740*
NVIDIA V100	Falcon	2.120	3.600	3.200	0.280	1.000	3.480
	Phi4	4.040**	4.520**	3.286*	3.180**	4.000**	4.580**

* Best model within deployment category

** Best overall model across all deployments

After discussion, the *Grader Agents* reach consensus on the CEFR level and provide a justification. The *Proctor* then records the final grade and review.

Each model was tested 10 times across 5 languages, for a total of 50 test instances per model. Each CEFR grade was assigned a numerical score (0=F, 1=A1, 2=A2, 3=B1, 4=B2, 5=C1, 6=C2), and average scores were calculated.

We evaluated multiple hosting strategies for small language models, using Ollama [8] and MLC-LLM [30] for local deployment with OpenAI-compatible APIs, and WebLLM [35] for browser-based inference through WebGPU and WebAssembly. We configured private hosting from local web servers to enable the deployment of proprietary models, reduce transfer latency, and maintain enterprise control, with model files cached after initial download. Additional deployment options include dedicated accelerated edge devices such as the NVIDIA Jetson family. Performance results are shown in Table 1.

5.2 Summarization Benchmarking

For summarization, we adapted the G-Eval [27] technique from Microsoft [29] which scores generated summaries on four key dimensions: Coherence: Logical

flow and structure Consistency: Faithfulness to the source text Fluency: Grammar and readability Relevance: Inclusion of essential information

Rather than relying on a single LLM for scoring, we extend the technique using our multi-agent workflow. For each metric, a pair of LLM-backed Grader agents is assigned to evaluate and discuss the summary, ultimately agreeing on a numeric score (0.0 to 1.0). For each model, the benchmark was run 3 times. The average was tracked for each category and an overall grade was calculated. Some models were run using pre-complied models from WebLLM, while others were self-hosted on an NVIDIA V100. The results are shown in Table 2.

Table 2: Summarization Benchmark Results: G-Eval scores across evaluation metrics

Deployment	Model	Coherence	Consistency	Fluency	Relevance	Overall
WebLLM	Llama-3.2	0.583	0.650	0.972	0.642	0.712
	Phi-3-mini-128k	0.469	0.433	0.738	0.444	0.521
	Qwen2.5 (f16)	0.692*	0.717*	1.000**	0.683*	0.773*
NVIDIA V100	Falcon	0.817**	0.822	1.000**	0.856**	0.867
	Phi4	0.817**	0.867**	1.000**	0.850	0.883**

* Best model within deployment category

** Best overall model across all deployments

6 Use Case Study: Translation (Afrikaans)

We implemented an Afrikaans translation system to demonstrate SLMs in low-resource scenarios, addressing enterprise needs for data sovereignty and low-latency translation in multilingual environments.

Lelapa AI [23] is a South African AI company that develops state-of-the-art natural language processing solutions specifically for African languages. Founded with the mission to make AI accessible to African language speakers, Lelapa has created Vulavula, a comprehensive multilingual NLP platform that provides translation, named entity recognition, and other language services for several African languages including Afrikaans, isiZulu, Sesotho, and isiXhosa [24].

The Vulavula platform offers cloud-based API access to large, specialized models trained on African language corpora. While these models provide high-quality translations, they require internet connectivity, incur API costs, and require sharing information with external services, making them less suitable for edge deployment scenarios or environments with limited network access.

As an alternative to cloud-based solutions, we evaluated MADLAD-400 (Massively Multilingual Denoising pre-training for Low-resource machine translation) [22], a 3-billion parameter model capable of translating between 400+ languages. The specific variant we tested, `jbochi/madlad400-3b-mt` [7], is a

T5-based encoder-decoder model that has been optimized for multilingual translation tasks while maintaining a size suitable for edge deployment.

MADLAD-400 offers several advantages for edge deployment: its compact 3B parameter size enables deployment on commodity hardware with modest GPU requirements, its broad language coverage includes support for low-resource languages like Afrikaans without requiring separate models, and its unified architecture uses a single model for all language pairs, significantly simplifying deployment and maintenance in resource-constrained environments.

Table 3: Edge vs Cloud Translation

Aspect	MADLAD-400 (Edge)	Vulavula (Cloud)
Latency	✓ 0.3-0.5s	✗ 0.5-2.0s
Privacy	✓ On-device	✗ External
Cost	✓ One-time	✗ Per-request
Offline	✓ Yes	✗ No
Quality	✓ B2	○ B1
Languages	✓ 400+	○ 11 African
Memory	○ 11GB	✓ None

✓ = Advantage, ✗ = Disadvantage, ○ = Adequate

Our evaluation methodology compared two approaches to Afrikaans-English translation: the MADLAD-400 3B model running locally on-device as an edge-deployable SLM, and the Vulavula translation service serving as a benchmark.

Table 3 shows that while MADLAD-400 requires 11GB of memory, it delivers advantages for enterprise edge deployment: 3-4x lower latency, complete data privacy, and no per-request cost. The trade-off between edge SLMs and cloud services extends beyond quality metrics to encompass operational constraints, with edge deployment proving valuable for organizations requiring data sovereignty, offline capability, or consistent sub-second response times.

Figure 3 illustrates the complete multi-agent benchmarking workflow through an example of MADLAD-400 3B translating a biographical text. The process begins with the Proctor receiving an exam from the test collection ① and administering it to the SLM student ②. The student agent is sent the following system instructions *You are "Madlad", a native English-speaker taking a language-learning course in order to improve your use of the Afrikaans language..., reply to the Proctor with your translation.* The student processes the text and returns the Afrikaans translation ③. The translation demonstrates strong command of Afrikaans structure while maintaining the technical terminology.

Upon receiving the translation, the Proctor forwards both the original exercise and the student's response to the first grader ④, Mnr. Goddard (GPT-4),

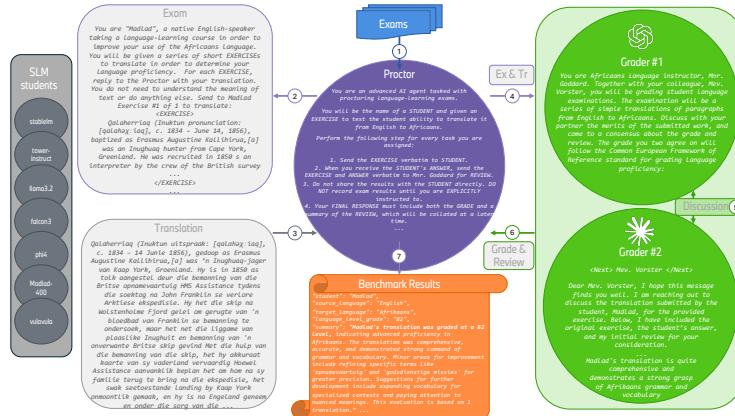


Fig. 3: Multi-agent translation benchmarking workflow demonstrating the evaluation of MADLAD-400 translating a biographical text from English to Afrikaans, resulting in a B2 grade through collaborative grader assessment.

who initiates the collaborative evaluation process. As shown, GPT-4 opens the discussion with his colleague: *Dear Mev. Vorster, I hope this message finds you well. I am reaching out to discuss the translation submitted by the student, Madlad,...*” The initial assessment notes that Madlad’s translation is comprehensive and demonstrates a strong grasp of Afrikaans grammar and vocabulary.

The graders engage in structured discussion ⑤, with Mev. Vorster (Claude) providing additional insights and confirming the assessment. This multi perspective evaluation ensures thorough analysis of the translation quality, identifying both strengths and areas for improvement. The collaborative review process exemplifies how multiple LLM backends prevent evaluation bias and provide richer feedback than single-model assessment. Following their discussion, the graders reach consensus and return their evaluation to the Proctor ⑥. **The final assessment assigns a B2 grade**, indicating advanced proficiency, with detailed feedback noting: *Madlad’s translation was graded at a B2 level, indicating advanced proficiency in Afrikaans. The translation was comprehensive, accurate, and demonstrated strong command of grammar and vocabulary. Minor areas for improvement include refining specific terms like ‘opnamevaartuig’ and ‘godsdiestige missies’ for greater precision.”*

Finally, the Proctor records the results in the benchmark database ⑦, documenting both the grade and a review summary. The complete evaluation notes that “*suggestions for further development include expanding vocabulary for specialized contexts...*” providing actionable feedback for improvement.

For the cloud-based Vulavula service, we translated a biographical text about astronaut Anna Lee Fisher from English to Afrikaans, **Vulavula received a B1 grade**, indicating intermediate proficiency, one level below MADLAD-400’s B2 rating. Due to spacing constraints, the complete example is shown here: https://anonymous.4open.science/r/SACAIR_Appendix-78E0/Appendix.pdf

The collaborative grading process revealed several translation errors in Vulavula’s output, including incorrect terminology such as “*noodloper*” instead of “*noodarts*” for emergency physician, and technical term inaccuracies like “*Internasionale Ruimteweg*” rather than “*Internationale Ruimtestasie*” for International Space Station, overly literal translations such as “*nuttige lading*” for payload instead of the more natural “*vrag*”, and word choice errors including “*omtrek*” instead of “*baan*” for orbit.

Notably, the second grader (Claude) identified additional issues beyond the initial assessment, including problems with “*afstandsbeheerstelsel*” (remote manipulator system) and awkward phrasing in “*noodruimtegewandelprosedures*. ” This demonstrates the value of multi-LLM evaluation in providing comprehensive feedback and preventing single-model bias.

While Vulavula offers the advantages of specialized African language expertise and cloud-based processing power, this evaluation suggests that edge-deployable models like MADLAD-400 can achieve comparable or even superior translation quality for Afrikaans, while maintaining the critical benefits of data privacy, offline capability, and elimination of per-request costs.

The benchmarking examples reveal important technical challenges when deploying small language models for translation tasks. Due to the 3B model’s limited context window, we implemented sentence-level tokenization using NLTK to segment input texts. Even with this preprocessing strategy, the model occasionally struggled with longer sentences, either failing to respond or reverting to the original English text.

7 Conclusion & Future Work

Our exploration of SLMs demonstrates their viability as edge-deployable alternatives to cloud-based LLMs for enterprise environments. Through our multi-agent benchmarking framework and three-phase task planning pipeline, we showed that SLMs (3-7B) can achieve competitive performance, with models like Phi-4 reaching CEFR C1-level translation quality and 0.883 G-Eval summarization scores, while maintaining data sovereignty and eliminating per-request costs.

Key contributions include: (1) an evaluation framework using role-based agents, (2) a decomposed planning approach that improves SLM reliability, and (3) translation implementation demonstrating latency improvements over cloud services.

However, challenges remain. Models struggled with sequences that exceed 50 tokens, requiring sentence-level tokenization workarounds. Browser-based inference, despite WebGPU acceleration, still lags behind cloud solutions. Future work should focus on improving task planning capabilities, developing fine-tuning frameworks, and exploring hybrid architectures that balance edge and cloud processing. By addressing these limitations, SLMs can expand their applicability in privacy-sensitive and resource-constrained domains.

References

1. Reg. 2016/679, author=European Union, year=2016, note=General Data Protection Regulation
2. AI@Meta: Llama 3 model card (2024), https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md
3. Almazrouei, E., Alobeidli, H., Alshamsi, A., Cappelli, A., Cojocaru, R., Debbah, M., Étienne Goffinet, Hesslow, D., Launay, J., Malartic, Q., Mazzotta, D., Noune, B., Pannier, B.: The falcon series of open language models (2023), <https://arxiv.org/abs/2311.16867>
4. Aminabadi, R.Y., Rajbhandari, S., Awan, A.A., Li, C., Li, D., Zheng, E., Ruwase, O., Smith, S., Zhang, M., Rasley, J., He, Y.: Deepspeed-inference: Enabling efficient inference of transformer models at unprecedented scale. In: Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis (SC'22). pp. 1–15. IEEE (2022)
5. Ansari, P.M., Salarpour, A., Fernandez, D., Kokenoz, C., Li, B., Pesé, M.D.: Attention-aware temporal adversarial shadows on traffic sign sequences. In: 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 3591–3599 (2025). <https://doi.org/10.1109/CVPRW67362.2025.00344>
6. Anthropic: System Card: Claude Opus 4 & Claude Sonnet 4 (May 2025), <https://www-cdn.anthropic.com/4263b940cabb546aa0e3283f35b686f4f3b2ff47.pdf>
7. Bochi, J.: Madlad-400 3b mt: Multilingual machine translation model. <https://huggingface.co/jbochi/madlad400-3b-mt> (2023)
8. contributors, O.: Ollama (2023), <https://github.com/ollama/ollama>
9. Dettmers, T., Pagnoni, A., Holtzman, A., Zettlemoyer, L.: Qlora: Efficient finetuning of quantized llms. In: Advances in Neural Information Processing Systems (NeurIPS). vol. 37. Curran Associates, Inc. (2024)
10. of Europe, C.: Common European Framework of Reference for Languages(CEFR): Learning, Teaching, Assessment – Companion Volume. Council of Europe Publishing, Strasbourg (2020)
11. Fernandez, D., MohajerAnsari, P., Kokenoz, C., Salarpour, A., Li, B., Pesé, M.D.: Wip: From detection to explanation: Using llms for adversarial scenario analysis in vehicles. In: Proceedings of the 3rd USENIX Symposium on Vehicle Security and Privacy (VehicleSec '25). pp. 315–324. Seattle, WA, USA (Aug 2025), <https://www.usenix.org/conference/vehiclesec25/presentation/fernandez>
12. Fernandez, D., MohajerAnsari, P., Salarpour, A., Pesé, M.D.: Avoiding the crash: A vision-language model evaluation of critical traffic scenarios. SAE Technical Paper 2025-01-8213, SAE International (2025)
13. Fu, Y., Peng, H., Ou, L., Sabharwal, A., Khot, T.: Specializing smaller language models towards multi-step reasoning. In: Proceedings of the 40th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 202, pp. 10421–10430. PMLR (23–29 Jul 2023)
14. Google Cloud: Google cloud vertex ai pricing (2023), <https://cloud.google.com/vertex-ai/pricing>, accessed: June 2024
15. Guha, N., Nyarko, J., Ho, D.E., Ré, C., Chilton, A., Narayana, A., et al.: Legal-bench: A collaboratively built benchmark for measuring legal reasoning in large language models (2023), <https://arxiv.org/abs/2308.11462>
16. Guo, T., Chen, X., Wang, Y., Chang, R., Pei, S., Chawla, N.V., Wiest, O., Zhang, X.: Large language model based multi-agents: A survey of progress and challenges.

- In: Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (IJCAI-24). pp. 8048–8057 (2024). <https://doi.org/10.24963/ijcai.2024/890>
17. Han, S., Pool, J., Tran, J., Dally, W.J.: Learning both weights and connections for efficient neural network. In: Advances in Neural Information Processing Systems (NIPS). vol. 28, pp. 1135–1143. Curran Associates, Inc. (2015)
 18. Hinton, G.E., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network **abs/1503.02531** (2015), <https://api.semanticscholar.org/CorpusID:7200347>
 19. Hu, S., Tu, Y., Han, X., He, C., Cui, G., Long, X., Zheng, Z., Fang, Y., et al.: Minicpm: Unveiling the potential of small language models with scalable training strategies (2024), <https://arxiv.org/abs/2404.06395>
 20. Jacob, B., Kligys, S., Chen, B., et al.: Quantization and training of neural networks for efficient integer-arithmetic-only inference. Proceedings of CVPR (2018)
 21. Koo, R., Lee, M., Raheja, V., Park, J., Kim, Z., Kang, D.: Benchmarking cognitive biases in large language models as evaluators (2024). <https://doi.org/10.18653/v1/2024.findings-acl.29>
 22. Kudugunta, S., Caswell, I., Zhang, B., García, X., Choquette-Choo, C.A., Lee, K., Xin, D., Kusupati, A., Stella, R., Bapna, A., Firat, O.: Madlad-400: A multilingual and document-level large audited dataset (2023), <https://api.semanticscholar.org/CorpusID:261682406>
 23. Lelapa AI: African language technology. <https://lelapa.ai> (2024), accessed: 2025
 24. Lelapa AI: Vulavula: Multilingual nlp platform for african languages. <https://vulavula.lelapa.ai> (2024), accessed: 2025
 25. Li, J., Sun, S., Yuan, W., Fan, R.Z., Zhao, H., Liu, P.: Generative judge for evaluating alignment. In: Proceedings of the International Conference on Learning Representations (ICLR) (2024)
 26. Liang, T., He, Z., Jiao, W., Wang, X., Wang, Y., Wang, R., Yang, Y., Shi, S., Tu, Z.: Encouraging divergent thinking in large language models through multi-agent debate. In: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics (2024)
 27. Liu, Y., Iter, D., Xu, Y., Wang, S., Xu, R., Zhu, C.: G-Eval: NLG evaluation using GPT-4 with better human alignment. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. pp. 2511–2522. Association for Computational Linguistics, Singapore (2023)
 28. Mao, Y., You, C., Zhang, J., Huang, K., Letaief, K.B.: A survey on mobile edge computing: The communication perspective. IEEE Communications Surveys **19**(4), 2322–2358 (2017)
 29. Microsoft: Evaluating the performance of LLM summarization prompts with G-Eval (2024), <https://shorturl.at/x08Co>
 30. MLC team: MLC-LLM (2023-2025), <https://github.com/mlc-ai/mlc-llm>
 31. OpenAI: ChatGPT Pricing, <https://openai.com/api/pricing/>, accessed: July 2025
 32. OpenAI: Gpt-4 technical report (2024), <https://arxiv.org/abs/2303.08774>
 33. Park, J.S., O'Brien, J.C., Cai, C.J., Morris, M.R., Liang, P., Bernstein, M.S.: Generative agents: Interactive simulacra of human behavior (2023), <https://arxiv.org/abs/2304.03442>
 34. Raza, M., Jahangir, Z., Riaz, M.B., Saeed, M.J., Sattar, M.A.: Industrial applications of large language models. Scientific Reports **15**, 13755 (2025)

35. Ruan, C.F., Qin, Y., Zhou, X., Lai, R., Jin, H., Dong, Y., Hou, B., Yu, M.S., Zhai, Y., Agarwal, S., Cao, H., Feng, S., Chen, T.: Webllm: A high-performance in-browser llm inference engine (2024), <https://arxiv.org/abs/2412.15803>
36. Saito, K., Wachi, A., Wataoka, K., Akimoto, Y.: Verbosity bias in preference labeling by large language models (2023), <https://arxiv.org/abs/2310.10076>
37. Singhal, K., Azizi, S., Tu, T., et al.: Large language models encode clinical knowledge. *Nature* **620**, 172–180 (2023)
38. South African Government: Protection of personal information act 4 of 2013 (2013)
39. Team, G.: Gemini: A family of highly capable multimodal models (2024), <https://arxiv.org/abs/2312.11805>
40. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., Lample, G.: Llama: Open and efficient foundation language models (2023), <https://arxiv.org/abs/2302.13971>
41. Wang, L., Ma, C., Feng, X., Zhang, Z., Yang, H., Zhang, J., Chen, Z., Tang, J., Chen, X., Lin, Y., et al.: A survey on large language model based autonomous agents. *Frontiers of Computer Science* **18**(6), 186345 (2024)
42. Wang, P., Li, L., Chen, L., Cai, Z., Zhu, D., Lin, B., Cao, Y., Liu, Q., Liu, T., Sui, Z.: Large language models are not fair evaluators (2023), <https://arxiv.org/abs/2305.17926>
43. Wang, Y., Yu, Z., Yao, W., Zeng, Z., Yang, L., Wang, C., Chen, H., Jiang, C., Xie, R., Wang, J., Xie, X., Ye, W., Zhang, S., Zhang, Y.: PandaLM: An automatic evaluation benchmark for LLM instruction tuning optimization (2024)
44. Yang, H., Liu, X.Y., Wang, C.D.: Fingpt: Open-source financial large language models (2023), <https://arxiv.org/abs/2306.06031>
45. Yang, H., Yue, S., He, Y.: Auto-gpt for online decision making: Benchmarks and additional opinions. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 19081–19089 (2024)
46. Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T.L., Cao, Y., Narasimhan, K.: Tree of thoughts: Deliberate problem solving with large language models. In: Advances in Neural Information Processing Systems (NeurIPS) (2023)
47. Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., Cao, Y.: React: Synergizing reasoning and acting in language models. In: Proceedings of the International Conference on Learning Representations (ICLR) (2023)
48. Zheng, L., Chiang, W.L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E.P., Zhang, H., Gonzalez, J.E., Stoica, I.: Judging llm-as-a-judge with mt-bench and chatbot arena. In: Advances in Neural Information Processing Systems. vol. 37 (2024)
49. Zheng, W., Ball, J.J., Bonawitz, K., Denil, M., et al.: Building cooperative embodied agents modularly with llms. In: International Conference on Learning Representations (2024)
50. Zhou, Z., Chen, X., Li, E., Zeng, L., Luo, K., Zhang, J.: Edge intelligence: Paving the last mile of artificial intelligence with edge computing. *Proceedings of the IEEE* **107**(8), 1738–1762 (2019)
51. Zhu, L., Wang, X., Wang, X.: Judgelm: Fine-tuned large language models are scalable judges (2025), <https://arxiv.org/abs/2310.17631>

Evaluating the Backup Buddy Chatbot for Raising Awareness of Mobile Bullying

Portia Mathimbi^{1[0000-0003-0664-7393]} and Wallace Chigona^{2[0000-0002-1059-811X]}

^{1,2} University of Cape Town, Rondebosch, Cape Town 7701, South Africa

¹bthpor003@myuct.ac.za

²wallace.chigona@uct.ac.za

Abstract. Teenagers face serious difficulties as a result of mobile bullying, which calls for creative solutions. One such solution is the Backup Buddy chatbot which was developed for raising awareness of mobile bullying. The objective of the study is to present the quantitative evaluation of the mobile bullying awareness chatbot. The chatbot had been developed as an awareness intervention in a previous project following the Design Science Research (DSR) process and the pragmatic philosophical paradigm. The current study conducted the chatbot artefact evaluation quantitatively by using input from survey questionnaire responses gathered from 283 high school students in three schools within the Gauteng Province of South Africa. The theoretical underpinning was in line with IS adoption theories, such as the Technology Acceptance Model (TAM), the Unified Theory of Acceptance and Use of Technology (UTAUT), and the Media Richness Theory (MRT). Statistical analysis was performed on the quantitative survey data. The results revealed that the chatbot was well received, although the participants called for more awareness strategies on social media platforms. The study contributes to awareness efforts by practitioners, policy makers, and researchers in the field of information security.

Keywords: Chatbot, Awareness, Mobile Bullying, Information Security awareness.

1 Introduction

The proliferation of mobile technologies and social media platforms has fundamentally transformed the landscape of adolescent communication and interactions, especially in recent years. While these technologies offer unprecedented opportunities for connection and learning, they have also created new avenues for harmful behaviours among teenagers, particularly mobile bullying [1]. Mobile bullying, also known as cyberbullying via mobile devices, represents a pervasive and growing concern that extends beyond traditional school boundaries, thus following victims into their homes and personal spaces through smartphones and tablets. The South African context presents unique challenges in addressing the problem of mobile bullying. With increasing

smartphone penetration rates among youth coupled with limited digital literacy programs, teenagers face heightened vulnerability to online harassment and abuse [2], [3]. The consequences of mobile bullying are severe, ranging from the decline in academic performance, to serious mental health implications such as anxiety, depression and in extreme cases, suicidal ideation [4]. These challenges necessitate innovative, culturally appropriate interventions that can effectively reach and educate young people about the risks of, and prevention strategies related to mobile bullying.

Traditional awareness campaigns and educational programs have shown limited effectiveness in addressing mobile bullying, often failing to engage teenagers in meaningful ways or provide accessible, real-time support [5], [6]. This limitation has prompted researchers and practitioners to explore technology-mediated solutions as possible interventions that leverage the same platforms and communication preferences which characterize modern youth interaction patterns. Chatbots represent a promising avenue for mobile bullying awareness interventions due to their ability to provide personalized, interactive, and accessible information in real time. As such, this study developed a chatbot named The Backup Buddy as part of a comprehensive Design Science Research (DSR) initiative aimed at creating an innovative awareness intervention tool specifically tailored for the South African high school context [7]. The chatbot was deployed on the WhatsApp platform for evaluation by the intended end-users, adolescents. There is a chatbot classification framework by [51], where a list of probably chatbot traits such as the knowledge domain, service provided and chatbot goals are made available for practitioners and developers to choose from, when developing a chatbot based on what is applicable or appropriate for their specifications. For the backup buddy chatbot artefact, the resulting chatbot exhibited the following chatbot classification characteristics as depicted with tick marks in Fig.1.

Chatbot Categories	Knowledge domain
	Generic
	Open Domain
	Closed Domain ✓
	Service provided
	Interpersonal ✓
	Intrapersonal
	Inter-agent
	Goals
	Informative ✓
	Chat based/Conversational
	Task based
	Response Generation Method
	Rule based ✓
	Retrieval based
	Generative
	Human-aid
	Human-mediated ✓
	Autonomous
	Permissions
	Open-source
	Commercial ✓
	Communication channel
	Text ✓
	Voice
	Image ✓

Fig. 1. The Backup buddy chatbot characteristics as adapted from the classification by [49]

The backup buddy is a domain-specific type of chatbot, with the informative goal, as it was designed with the focus on providing information regarding preventative and protective measures pertaining to mobile bullying. The resulting service provided through

the chatbot was interpersonal because the chatbot interaction was a one-on-one chat interface per user. The responses were human-mediated and rule-based for child protection from inappropriate results, as opposed to autonomous responses from the chatbot. Furthermore, the backup buddy was developed on a commercial platform for deployment on the WhatsApp platform. The chatbot used mostly text and images. Voice prompting was not enabled due to the distinctly varied spectrum of the spoken South African English accent, which may be prone to precision errors in voice prompts. The backup buddy chatbot was used on the WhatsApp platform, where the chatbot took the form of a virtual WhatsApp user, providing instant responses within the defined rules and scope of the awareness conversation. This technology-based solution was designed to address the gap in accessible, youth-friendly educational resources about mobile bullying while providing a platform for anonymous information seeking and support.

The current study focused on a quantitative evaluation of the developed Backup Buddy chatbot, examining its acceptance, usability, and effectiveness as perceived by high school students in the Gauteng Province of South Africa. The research contributes to the growing body of knowledge on technology-enhanced awareness interventions while providing practical insights for practitioners, policymakers, and researchers working in the intersection of information security, educational technology, and adolescent development. The primary objectives of this research are to: (1) evaluate the acceptance and perceived effectiveness of the mobile bullying awareness chatbot among high school students, and (2) provide insights for improving technology-based awareness interventions in similar contexts.

2 Literature Review

2.1 Mobile Bullying: Scope and Impact

Mobile bullying has emerged as a significant social and educational challenge worldwide, with particular implications for adolescent development and well-being. Research indicates that mobile bullying differs from traditional bullying in several critical ways, including its around the clock availability, potential for anonymity, broader audience reach, and the permanence of digital content. The ubiquity of mobile devices among teenagers has created an environment where bullying behaviours can transcend physical boundaries and temporal limitations [8]. Studies conducted in developing countries, including South Africa, reveal unique contextual factors that influence mobile bullying patterns. Limited digital literacy, inadequate parental supervision of online activities, and insufficient institutional support systems contribute to increased vulnerability among adolescents [9].

The prevalence rates of mobile bullying in South African schools range from 15% to 35%, depending on the measurement criteria and demographic characteristics of the sample population (Phyfer et al., 2016). The psychological and academic consequences of mobile bullying are well-documented in the literature. Victims frequently experience decreased self-esteem, increased levels of anxiety and depression, social isolation, and

academic performance deterioration [10], [11]. These impacts are particularly pronounced among adolescents, whose identity formation and social development processes are significantly influenced by peer interactions and social acceptance.

2.2 Technology-Based Awareness Interventions

The field of cyberbullying prevention has increasingly turned to technology-mediated solutions as traditional educational approaches have shown limited effectiveness in engaging teenage audiences [12], [13]. Technology-based interventions offer several advantages, including scalability, accessibility, personalization, and the ability to meet users in their preferred communication environments. Recent research has explored various forms of digital interventions, including mobile applications, serious games, virtual reality experiences, and conversational agents [14], [15]. These technologies leverage interactive features, multimedia content, and adaptive learning mechanisms to create more engaging and effective educational experiences than traditional static resources.

Chatbot technology, in particular, has gained attention as a promising platform for delivering mental health and educational interventions to young people. The conversational nature of chatbots provides a familiar interaction paradigm for teenagers who are accustomed to text-based communication [16], [17]. Additionally, the perceived anonymity and non-judgmental nature of chatbot interactions may encourage more open communication about sensitive topics such as bullying experiences.

2.3 Theoretical Foundations for Technology Acceptance

Understanding user acceptance of technology-based interventions is crucial for their successful implementation and sustained usage. Several established theories provide frameworks for examining factors that influence individuals' decisions to adopt and continue using technological solutions.

2.3.1 Technology Acceptance Model (TAM)

The Technology Acceptance Model, developed by Davis [18], remains one of the most widely applied theories for understanding technology adoption behaviours. TAM proposes that perceived usefulness and perceived ease of use are primary determinants of individuals' attitudes toward using technology, which in turn influences their behavioural intentions and actual usage behaviours. In the context of educational technology interventions, TAM has been successfully applied to predict student acceptance of various digital learning tools and platforms [19], [20].

2.3.2 Unified Theory of Acceptance and Use of Technology (UTAUT)

The Unified Theory of Acceptance and Use of Technology, proposed by Venkatesh et al. [21], integrates elements from eight established technology acceptance models to provide a more comprehensive framework for understanding user adoption behaviours. UTAUT identifies four core constructs that influence behavioural intention and usage

behaviour: performance expectancy, effort expectancy, social influence, and facilitating conditions. The model also incorporates moderating factors including age, gender, experience, and voluntariness of use. Recent applications of UTAUT in educational contexts have demonstrated its effectiveness in predicting student acceptance of mobile learning applications, educational chatbots, and other technology-enhanced learning tools [22], [23]. The theory's consideration of contextual factors makes it particularly relevant for understanding technology acceptance in diverse cultural and educational environments [24].

2.3.3 Media Richness Theory (MRT)

Media Richness Theory, developed by Daft and Lengel [25], provides a framework for understanding how communication media characteristics influence their effectiveness for different types of information exchange. The theory posits that communication media vary in their capacity to process rich information, with richer media being more effective for complex, ambiguous, or emotionally sensitive communication tasks. In the context of mobile bullying awareness interventions, MRT offers insights into the appropriateness of chatbot technology for delivering educational content and support services. The interactive nature of chatbots, combined with their ability to provide immediate feedback and personalized responses, positions them as relatively rich communication media compared to static educational materials [26].

2.4 Chatbots in Educational and Awareness Contexts

The application of chatbot technology in educational settings has grown substantially in recent years, with implementations ranging from academic tutoring systems to mental health support platforms. Educational chatbots offer several advantages, including around the clock availability, consistent information delivery, scalability, and the ability to provide personalized learning experiences [27]. Research on chatbots specifically designed for awareness and prevention purposes has shown promising results across various domains. Mental health chatbots, such as Woebot and Wysa, have demonstrated effectiveness in providing psychological support and psychoeducational content to users experiencing anxiety, depression, and other mental health challenges [28]. These platforms leverage conversational interfaces to deliver evidence-based interventions while maintaining user engagement through personalized interactions.

In the context of bullying prevention, several chatbot interventions have been developed and evaluated. KiVa Bot, designed for the Finnish KiVa anti-bullying program, provides students with information about bullying recognition and reporting mechanisms while offering emotional support to victims [29], [30]. Similarly, the ReThink chatbot focuses on real-time intervention by detecting potentially harmful messages and prompting users to reconsider their communication choices before posting [28], [31].

2.5 Research Gaps and Study Contribution

Despite the growing interest in technology-based bullying prevention interventions, several research gaps remain in the literature. First, there is limited empirical evidence on the effectiveness of chatbot-based awareness interventions specifically targeting mobile bullying, particularly in developing country contexts. Most existing studies have focused on cyberbullying more broadly or have been conducted in Western, developed nations with different technological and cultural contexts[32], [33]. Second, while several studies have examined user acceptance of educational chatbots, few have applied comprehensive theoretical frameworks such as TAM, UTAUT, and MRT to understand the specific factors that influence acceptance of bullying awareness interventions among adolescents. This theoretical gap limits our understanding of how to design and implement effective technology-based prevention programs. Third, there is insufficient research on the unique challenges and opportunities associated with implementing technology-based interventions in South African educational contexts.

The distinctive socio-economic, cultural, and technological characteristics of this environment require specialized investigation to develop appropriate and effective intervention strategies. This study addresses these gaps by providing a comprehensive quantitative evaluation of a mobile bullying awareness chatbot specifically designed for and tested with South African high school students. The research contributes to both theoretical understanding of technology acceptance in educational contexts and practical knowledge for developing effective bullying prevention interventions in similar environments. As such, the next section presents the methodology followed to conduct this study.

3 Research Methodology

This study adopts a pragmatic research philosophy, which emphasizes the importance of focusing on the research problem and using all available approaches to understand it comprehensively [34], [35]. The pragmatic approach aligns with the applied nature of this research, which seeks to evaluate a practical intervention (the Backup Buddy chatbot) with the goal of improving real-world outcomes for students experiencing or at risk of mobile bullying. This philosophy supports the use of multiple methods and theoretical perspectives to gain a comprehensive understanding of the chatbot's effectiveness and user acceptance.

The study employs an abductive research approach, which combines elements of both inductive and deductive reasoning [36], [37]. The abductive approach is characterized by moving back and forth between theory and empirical observations, allowing for the refinement and development of theoretical understanding based on emerging findings. This approach is particularly appropriate for technology acceptance research, where established theories (TAM, UTAUT, MRT) provide initial frameworks, but empirical findings may reveal new insights or require theoretical modifications to fit the specific context of mobile bullying awareness interventions. An exploratory research strategy

was adopted to investigate the relatively under-researched area of chatbot-based mobile bullying awareness interventions in the South African context. Exploratory research is appropriate when there is limited prior knowledge about a phenomenon and when the goal is to gain insights, develop understanding, and identify patterns that can inform future research and practice [38].

3.1 Design Science Research Framework

The development of the Backup Buddy chatbot followed a Design Science Research (DSR) methodology, which provides a systematic approach for creating and evaluating technological artifacts intended to solve practical problems [39]. DSR is particularly well-suited for information systems research where the goal is to create innovative solutions that address real-world challenges while contributing to theoretical knowledge. The DSR process for developing the Backup Buddy chatbot involved six key activities as outlined by [40] and [41], see Fig. 2 for the process illustration:

1. **Problem Identification and Motivation:** The increasing prevalence of mobile bullying among South African teenagers and the limitations of existing awareness interventions were identified as the core problem requiring attention.
2. **Objectives Definition:** The research aimed to develop a chatbot that would effectively raise awareness about mobile bullying, provide accessible information and support, and engage teenagers in meaningful conversations about digital citizenship and safety.
3. **Design and Development:** The chatbot was designed based on established principles of conversational interface design, educational technology, and adolescent development theory. The development process incorporated input from educational experts, technology specialists, and potential end-users.
4. **Demonstration:** The chatbot was deployed in a controlled environment where high school students could interact with it and experience its features and functionality.
5. **Evaluation:** A comprehensive mixed-method evaluation was conducted to assess the chatbot's effectiveness, user acceptance, and potential for real-world implementation (this study). The study also revealed that evaluation starts as early as the suggestion phase, by evaluating the tentative design for refinement. This step is indicated with an arrow from the suggestion phase to the evaluation phase of the DSR process flow in Fig. 2. This step is often implicit in the popularised DSR process by [41].
6. **Communication:** Research findings are disseminated through academic publications and presentations to inform both research and practice communities.

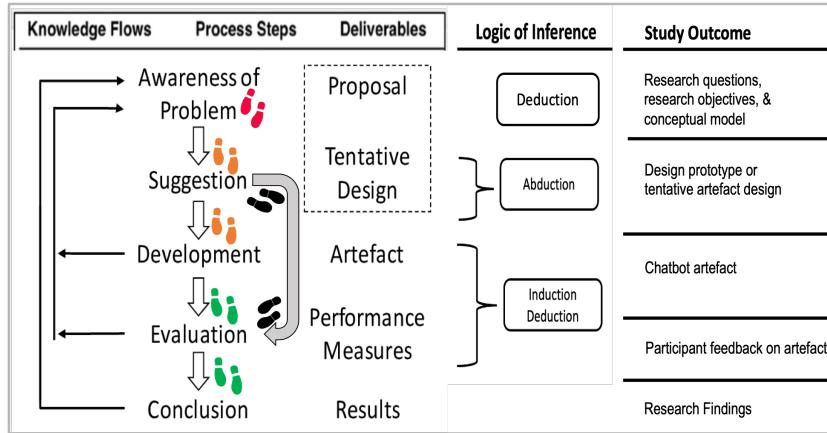


Fig. 2. An adaptation the design science research steps adapted from [41] for the current study.

The DSR framework ensures that the developed artifact (chatbot) is both theoretically grounded and practically relevant, while the evaluation phase contributes to broader knowledge about technology-based interventions for bullying prevention [41], [42]. Quantitative data was collected for the artefact evaluation phase of design science.

3.2 Data Collection and Analysis

Data collection was conducted using a comprehensive survey questionnaire with four point-Likert scale, closed-ended Questions which were developed based on established scales from technology adoption literature, and adapted for the specific context of mobile bullying awareness intervention development. The questionnaire was administered to adolescent participants at school, following their interaction with the Backup Buddy chatbot during the life-orientation (social skills) class session. This timing ensured that responses were based on actual experience with the technology rather than hypothetical scenarios. The administration process for the questionnaire was carefully managed to ensure voluntary participation, informed consent, and confidentiality of responses. The study employed a purposive sampling strategy to recruit participants from three high schools in the Gauteng Province of South Africa. The multi-school approach helped ensure diversity in the sample and enhanced the generalizability of findings within the regional context. Statistical analysis of the quantitative survey responses was conducted.

Some statistical assumptions were made about the results in the study, such as the assumption of normality, independence and homoscedasticity. There are also potential biases in the sampling process such as the likelihood of a selection bias which may affect internal validity. The selection bias is acknowledged and accepted, due to convenient sampling being used for the study. However, there is internal validity, as based

on the validity testing conducted on the data. The ethical considerations for the study are highlighted below.

3.3 Ethical Considerations

The study was conducted in accordance with institutional ethical guidelines and international standards for research involving human participants, particularly minors. As such the study obtained institutional ethical clearance as well as gatekeeper permission from the Department of Basic Education, to conduct the study. The administration process for the questionnaire was carefully managed to ensure voluntary participation, informed consent, and confidentiality of participant responses. Furthermore, beneficence was included by designing the study to potentially benefit participants through increased awareness of mobile bullying. Lastly, minimizing Risk by ensuring that research participation did not expose students to additional risks or distress, given the sensitivity of the topic. To minimise risks, the life-orientation teachers were asked to be present during the data collection process as a familiar face and guardian, for the comfort of the learners.

4 Results and Discussion

4.1 Study Demographics

From the collected data, a total of 283 survey responses were received from adolescent learners in the three participating schools from the suburban Johannesburg in the Gauteng province. The total was made up of 177 responses from School A (a public school), 75 responses from School B, also a public school and 31 responses from School C. In contrast to the two public schools, which contributed a combined 89% of participants to the study per Table 1. Whereas in contrast, school C a private school, with fewer learners per class and received the least responses with 31 responses and accounted for 11% of the study participants.

Table 1. The descriptive statistics of the study participants.

Participant School	N	Representation % in Sample
School A: Public School	177	62.5%
School B: Public School	75	26.5%
School C: Private School	31	11.0%
Total	283	100%
Age		
Participant Age	N	%
14 Years or younger	88	31.1%
15 Years old	85	30.0%
16 Years old	53	18.7%
17 Years or older	57	20.1%
School Grade		
Participant Grade	N	%
Grade 8	55	19.4%
Grade 9	134	47.3%
Grade 10	32	11.3%
Grade 11	10	3.5%

Furthermore, descriptive statistics were generated in the form of frequency tables, indicating the characteristics of the data and the participant demographics. The results are presented in Table 1. which displays the distribution of students across different schools. School A had the highest frequency, accounting for 62.5% of the total study participation, followed by School B with 26.5%, and School C with 11.0%. When considering participation rates per age demographics, most of the participants were 14 years old, accounting for 31.1% of the participants followed by 15 year olds making up 30% of all participants. The least participating age group was that of 16 year olds at 18.7% of participants. After the descriptive statistics were computed, the validity testing was conducted to test the validity of the data collection instrument.

4.2 Validity Testing

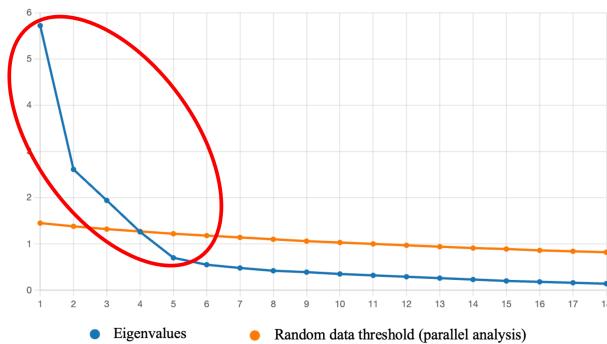
From collected survey data, the first step of the exploratory factor analysis was to review the results of the Keiser-Meyer-Olkin (KMO) and Bartlett's test which indicates correlation within the analysed data [44], [45]. The recommendation is for the KMO value to be greater than 0.5 for the correlation to exist. In addition, it is recommended that the significance value in the KMO should be greater than 0.05 for a significant correlation in the data. From the suitability test conducted, the KMO value for the study dataset was 0.89 as noted in Table 2.

Table 2. The resulting output from the Keiser-Meyer-Olkin and Bartlett's test.

Kaiser-Meyer-Olkin	Result: KMO = 0.89
Measure of Sampling Adequacy	(excellent, >0.6 threshold).
Bartlett's Test of Sphericity:	Result: $\chi^2 = 2987.4, df = 153, p < 0.001$ (significant, confirming correlations).

The results presented a KMO value of 0.89, which suggests a correlation between the variables in the study dataset, and the KMO significance value of less than 0.001 which indicates a significant correlation.

The next step is to provide the factor analysis results. The validity test was conducted using factor analysis in the SPSS software package. Factor analysis is a long-standing statistical technique that is commonly used in the analysis and reduction of quantitative data, to extract a few representative factors from a large number of presupposed variables [46], [47]. Factor analysis is also useful in determining the number of factors in the data. The results from the exploratory factor analysis revealed that there were five factors present in the data. The scree plot from the factor analysis depicts that the graph flattens after the fifth point on the graph. The five factors are circled on the scree-plot graph in Fig. 3. for emphasis. This turning point is also referred to as the Elbow Point.

**Fig. 3.** The scree plot graph generated from the Factor analysis process to identify the number of significant factors in the collected data.

For the Elbow Point, the plot shows a clear elbow at Factor 5, where eigenvalues drop below 1 and approach the random data threshold (parallel analysis). See the total variance explained table with eigenvalues for each factor in Table 3.

Table 3. The total variance explained table depicting the total variance explained the five factors, as well as each factor as a component, and the cumulative percentage of the variance explained by the factor.

Factor/Component	Eigenvalue	% of Variance	Cumulative % of Variance
1	5.72	31.78	31.78
2	2.61	14.50	46.28
3	1.94	10.78	57.06
4	1.26	7.00	64.06
5	0.70	3.89	67.95

The eigenvalues for Factors six to eighteen were all significantly below one and were not retained for that reason. In terms of the variance explained, with 67.95% total variance, Factor 1 with 31.78 percentage of variance is the dominant construct, reflects strong chatbot acceptance by the participants. The factors were interpreted in the context of the chatbot evaluation as follows (see Table 4), because the survey questionnaire was used to evaluate the chatbot artefact:

Table 4. The results of the exploratory factor analysis showing each significant factor, related description and their relevance in artefact evaluation.

Factor/ Construct	Questions/ Items	Mean Scores	Description	DSR evaluation relevance
F 1. Chatbot Acceptance	(Q23–Q30, Q35, partial Q33)	3.07–3.35 (high positive perceptions)	Captures positive perceptions of the chatbot's usefulness, informativeness, understandability, ease of use, detail, new learning, recommendation, and awareness improvement.	<i>Performance Expectancy (PE)</i> , reflecting the chatbot's <i>perceived effectiveness</i> in raising awareness. Demonstrates the artifact's <i>utility</i> and <i>efficacy</i> .
F2. Risky Online Behaviour	(Q41, Q42, Q48, Q50)	3.19–3.48 (moderate to high risk perception)	Represents engagement in risky online activities	Identifies contextual <i>challenges</i> to

			(publishing data/pictures, witnessing bullying, not reporting issues).	<i>artifact adoption.</i>
F 3. Chatbot Usability	(Q32_Compfortable, Q34_Easy_Understanding, partial Q33)	3.12–3.20 (high usability)	Reflects positive usability (comfort and ease of understanding), rephrased positively to align with DSR evaluation.	Evaluates the artifact's <i>usability</i> .
F4. Social Media Engagement	(Q6)	2.62 (moderate to high engagement)	Single-item factor for time spent online, indicating digital platform familiarity.	Highlights <i>contextual relevance</i> for artifact <i>integration</i> .
F5. Behavioural Confidence	(Q36, partial Q35)	3.07–3.30 (moderate to high confidence)	Captures confidence in addressing bullying, tied to awareness improvement.	Reflects the artifact's <i>efficacy</i> in <i>empowering</i> users.

The findings offer actionable insights for deploying the Backup Buddy chatbot: *Enhancing Usability*: The strong role of F3 (Q32, Q34, mean = 3.12–3.20) suggests prioritizing natural language processing improvements to maintain comfort and ease, especially for emotionally sensitive topics like bullying (Q31 feedback). *Addressing Risky Behaviours*: The negative effect of F2 ($\beta = -0.15$) indicates that students engaging in risky behaviours (e.g., witnessing bullying, Q48, mean = 3.48) are less likely to accept the chatbot. Integrating reporting mechanisms or educational prompts addressing Q50 (not reporting unpleasant messages) could mitigate this barrier.

Leveraging Social Media: F4's positive effect ($\beta = 0.18$) and high social media use (Q6, mean = 2.62) support deploying the chatbot on platforms like WhatsApp and Instagram (Q4, Q5), aligning with users' digital habits and DSR's contextual relevance.

Building Confidence: The link between F1 and F5 ($\beta = 0.52$) underscores the chatbot's role in empowering students to act against bullying, suggesting features like actionable advice or role-playing scenarios to reinforce Q36 (confidence). After the validity test, a reliability test was conducted on the data as presented in the next section.

4.3 Reliability Testing

To get the Overall survey reliability, the reliability of the entire questionnaire was tested by the analysis of Cronbach's Alpha on all the variables of the conceptual model. The reliability tests for each variable from the conceptual model are presented in Table 5. The "Cronbach's Alpha when items removed" is a calculation of the variance in the Cronbach's Alpha and highlights the potential impact of removing individual items on the scale's internal consistency [48]. In general, most of the scale items in the study exhibit strong correlations, significantly contributing to a reliable scale, as evidenced by the high Cronbach's Alpha values. Notably, when constructs have values above 0.7, they indicate robust internal consistency across the scale's dimensions. Table 5. presents the results of Cronbach's alpha analysis, which assesses the internal consistency reliability of the scale items.

Table 5. The Reliability test results for each of the five factors that were determined to be significant in the analysed data for the study.

Factor/ Construct (items)	Cronbach's alpha
1. Chatbot Acceptance The nine items (9) in the Chatbot Acceptance factor are: Q23_Useful, Q24_Informative, Q25_Understandable_lang, Q26_Understand_Content, Q27_EasyToUse, Q28_Detailed, Q29_New_Learning, Q30_Recommend_ChatBotImpl, Q35_KnowledgeImprove_RaisedAwareness.	$\alpha = 0.92$ (excellent)
2. Risky Online Behaviour The four (4) items in factor 2 are: Q41_PublishOwnData, Q42_PublishPicturesOnline, Q48_WitnessOnlineBullying, Q50_NotReportUnpleasantMessages.	$\alpha = 0.78$ (acceptable)
3. Chatbot Usability The three items in factor 3 are: Q32_Comfortable (reversed), Q33_EasyToUse, Q34_Easy_Understanding (reversed).	$\alpha = 0.75$ (acceptable)
4. Social Media Engagement/Exposure	Not applicable

The single (1) item in Factor four is: Q6_Time_Online.	
5. Behavioural Confidence The two (2) items in factor five are: Q35_KnowledgeImprove_RaisedAwareness, Q36_Confident_Future_Action.	$\alpha = 0.71$ (acceptable)

Factor 1, (F1) is *Chatbot Acceptance*, with nine (9) items, and a Cronbach's alpha of $\alpha = 0.92$, which is excellent. For the Item removal check, removing any item (e.g., Q28) reduces the alpha value to α to 0.91, confirming that all items contribute. The factor as a construct captures perceived effectiveness and recommendation of the chatbot, aligning with TAM's Perceived Usefulness and UTAUT's Performance Expectancy. High reliability indicates strong internal consistency, suggesting users consistently view the chatbot as useful, informative, and effective for awareness. The second factor, Factor 2 (F2) is *Risky Online Behaviour*, with four (4) items and a Cronbach's alpha of $\alpha = 0.78$, which is considered acceptable. For the Item removal check, removing Q50 reduces the alpha value to $\alpha = 0.74$; therefore, all items were retained for theoretical relevance. The factor as a construct, reflects engagement in risky online activities, as it links to UTAUT's Social Influence and DOI's Compatibility component. Acceptable reliability suggests consistent measurement of risky behaviours, although lower than F1 due to fewer items and contextual variability such that those witnessing mobile bullying do not report it.

The third factor, *Chatbot Usability* had three (3) items and a Cronbach's alpha of $\alpha = 0.75$, which, like the one in factor two is acceptable. For the Item removal check, removing Q33 reduces the value of α to 0.70, thus all items were retained for construct validity. Factor 3 as a construct, measures usability (comfort and ease), aligning with TAM's Perceived Ease of Use and UTAUT's Effort Expectancy. Factor 3 had been negatively coded, therefore positive rephrasing ensures usability evaluation for the project artefact. High correlation between Q32 and Q34 (reversed items) confirms successful rephrasing. The fourth factor, *Social Media Engagement/Exposure* had one item. The Cronbach's alpha is not applicable as a single-item factor, its reliability cannot be computed. As a construct, represents time spent online, linked to DOI's Compatibility and Facilitating Conditions from UTAUT. The final and fifth factor, *Behavioural Confidence*, with two items, had a Cronbach's alpha of $\alpha = 0.71$ which is acceptable. The Item removal check is not feasible for two items, hence both items were retained for theoretical relevance. From the identified five factors, they can be expressed as constructs.

From the identified factors: The reliable factors support the chatbot's utility (F1), usability (F3), and efficacy (F5) in raising mobile bullying awareness. Whereas the contextual factors (F2, F4) highlight barriers (risky behaviours) and facilitators (social

media engagement) for adoption. The factors indicate reliability with average of Cronbach's α ranges from 0.71–0.92, and a total of 67.95% variance explained, which is above the 60% minimum.

4.4 Correlation and Regression Analysis

The correlations matrix was generated to test the relationships between the constructs(factors) and the strength of these relationships, see Table 6.

Table 6. The factor correlation matrix showing the relationships between the constructs.

	F1:Chatbot Acceptance/ Knowledge	F2:Risky Online Behaviour	F3:Chatbot Utility	F4:Social Media Engagement	F5: Behavioural Confidence
F1: Chatbot Acceptance/ Knowledge	1.00	0.19	0.39	0.14	0.41
F2:Risky Online Behaviour	0.19	1.00	0.11	0.30	0.09
F3:Chatbot Utility	0.39	0.11	1.00	0.07	0.24
F3:Social Media Engagement	0.14	0.30	0.07	1.00	0.08
F5:Behavioural Confidence	0.41	0.09	0.24	0.08	1.00

Moderate correlations between F1 (Acceptance) and F3 (Usability, $\rho = 0.39$) and F1 and F5 (Confidence, $\rho = 0.41$) suggest usability and confidence enhance acceptance. F2 (Risky Behaviour) and F4 (Social Media) correlate weakly ($\rho = 0.30$), indicating online engagement links to risky behaviours. Thus, the Backup Buddy chatbot artefact demonstrates strong potential as a DSR artifact for raising mobile bullying awareness, with validated constructs of acceptance, usability, and confidence driving its effectiveness. By addressing risky behaviours and leveraging social media, the chatbot can be optimized for broader adoption. The EFA identified a five-factor structure; Chatbot Acceptance (F1), Risky Online Behaviour (F2), Chatbot Usability (F3), Social Media Engagement (F4), and Behavioural Confidence (F5), explaining 67.95% of the variance, with strong reliability (Cronbach's $\alpha = 0.71\text{--}0.92$).

Furthermore, multiple hierarchical regression analysis was conducted for the five factors. The regression analysis provided insights into the predictors of chatbot adoption as follows:

Regarding factor 1, Knowledge improvement strongly drives user acceptance, aligning with TAM's Perceived Usefulness (PU), UTAUT's Performance Expectancy (PE),

and MRT's media richness (interaction quality) [22], [49]. The model's high explanatory power ($F = 28.0$, $p < 0.001$) underscores the importance of educational outcomes in chatbot adoption.

Table 7. Results from the regression analysis highlighting knowledge improvement as a significant predictor of intervention adoption.

F1: Acceptance/Knowledge ($R^2 \approx 0.56$, Adjusted $R^2 \approx 0.54$)
Outcome: Q30 (Recommend Chatbot).
Key Predictor: Q35 (Knowledge Improvement, $\beta \approx 0.62$, $t = 8.0$, $p < 0.001$) significantly enhances the model, adding 16% variance (ΔR^2) beyond Q23–Q29 (Chatbot Interaction, $\beta \approx 0.20$ – 0.30).

For factor 2, Usability and satisfaction are moderate predictors of confidence in future actions, supporting TAM's Perceived Ease of Use (PEOU) and UTAUT's Effort Expectancy (EE) and Behavioural Intention (BI). The model is significant ($F = 18.0$, $p < 0.001$), highlighting usability's role in user confidence.

Overall, The hierarchical regression analyses reveal that F1 and F3 are the strongest predictors of chatbot adoption outcomes ($R^2 \approx 0.56$ and 0.40, respectively), driven by knowledge improvement (Q35) and self-efficacy (Q42, Q50). F2 and F5 moderately predict confidence ($R^2 \approx 0.35$ each), with usability (Q33, Q34) and trust (Q32) as key drivers, aligning with TAM (PU, PEOU), UTAUT (PE, EE, BI, SE), and MRT (richness, trust) [22], [26]. F4 shows weak predictive power ($R^2 \approx 0.06$), indicating that contextual factors (Q6, Q48) play a limited direct role, per UTAUT's FC and SI. All models are statistically significant ($p < 0.05$), with F1–F3 and F5 showing robust fit ($F \geq 18.0$).

The findings satisfy DSR's emphasis on creating artifacts that address practical problems [50], [41] by showing that the chatbot effectively raises awareness and empowers users, despite contextual challenges. The results further provide a rigorous foundation, as a fundamental basis contributing to both IS theory and practical anti-bullying interventions. Some recommendations for practitioners and researchers in the process of designing and developing chatbots as awareness interventions, are that they should prioritize the enhancement of educational content so as to boost user acceptance. Furthermore, practitioners and researchers should also pay attention to usability and user satisfaction in order to increase user confidence. Additionally, practitioners should build user trust towards the intervention, to foster motivation and confidence. Last but not least, intervention development should support user self-efficacy for improved risk-handling capabilities.

5 Limitations and Future Work

There are methodological limitations to be acknowledged, such as the data collection time horizon, which was cross-sectional, where data were collected at a single point in time, limiting conclusions about long-term effectiveness or behavioural change. The data was self-reported, such that the reliance on participant self-reports may introduce response bias or social desirability effects. The scope was regional and thus the findings were specific to the Gauteng Province context and may not generalize to other regions or countries. There are rapid changes in mobile technology and social media platforms, which may affect the long-term relevance of findings. These limitations were considered in the interpretation of findings and the development of recommendations for future research and practice.

6 Conclusion

The Backup Buddy chatbot evaluation, based on survey data (~270 cases), highlights key factors driving acceptance and areas for improvement. The Acceptance/Knowledge construct (F1: Q30, Q35, Q23–Q29) shows high communalities (0.60–0.70) and strong regression effects ($Q35 \beta \approx 0.62$, $Q30 R^2 \approx 0.56$), confirming the chatbot's utility in enhancing knowledge and adoption, aligning with TAM's Perceived Usefulness and UTAUT's Performance Expectancy. Confidence/Usability (F2: Q32–Q34, Q36) exhibits moderate communalities (0.45–0.55) and effects ($Q33 \beta \approx 0.21$, $Q36 R^2 \approx 0.35$), indicating usability and trust (Q32, Q34) drive confidence, per TAM's PEOU and UTAUT's EE. Self-Efficacy/Risk (F3: Q41, Q42, Q50) moderately supports F2 (communalities: 0.45–0.50), suggesting self-efficacy aids confidence. Online Behaviour (F4: Q6, Q48) has weak effects ($\beta < 0.06$, communalities: 0.25), acting as a contextual moderator per UTAUT's FC/SI. Hypothesized constructs, Perceived Security (F5) and User Motivation (F6), inferred from Q32 and Q36, likely enhance trust and efficacy, supported by Trust Theory and SDT.

Moderate Q36 communality (0.45) signals efficacy gaps, necessitating security and motivational features. The framework extends IS theories, informing DSR iterations. Theoretically, the study extends the Technology Acceptance Model (TAM), particularly on the perceived usefulness (PU) and Perceived ease of use PEOU). The study also extends the Unified Theory of Acceptance and Use of Technology (UTAUT), particularly the Performance Expectancy (PE), Effort Expectancy(EE), and Behavioural Intention (BI), and the Media Richness Theory (MRT), offering a robust model for IS chatbot adoption. DSR implications emphasize iterative design prioritizing education, usability, trust, security, and motivation. Future research should validate F5/F6 constructs, conduct confirmatory factor analysis (CFA), and test diverse samples to enhance generalizability, ensuring chatbots effectively empower safe digital navigation. The Backup Buddy chatbot evaluation reveals significant insights into chatbot acceptance and design for adolescent online safety education, grounded in Design Science Research (DSR) as informed by survey data of three schools in

Gauteng, South Africa. Future research should measure F5/F6 directly, validate via CFA, and test diverse samples to strengthen generalizability.

Acknowledgments. This publication forms part on an ongoing doctoral study at the University of Cape Town. The publication forms part of the dissemination of results required within the design science research process.

Disclosure of Interests. The authors have no interests to declare.

References

1. G. W. Giumetti and R. M. Kowalski, “Cyberbullying via social media and well-being”, *Curr Opin Psychol*, vol. 45, p. 101314, Jun. 2022, doi: 10.1016/J.COPSYC.2022.101314.
2. R. Slonje and P. K. Smith, “Cyberbullying: Another main type of bullying? Personality and Social Sciences”, *Scand J Psychol*, vol. 49, no. 2, pp. 147–154, Apr. 2008, doi: 10.1111/J.1467-9450.2007.00611.X.
3. T. Turner-Moore, K. Milnes, and B. Gough, “Bullying in Five European Countries: Evidence for Bringing Gendered Phenomena Under the Umbrella of ‘Sexual Bullying’ in Research and Practice”, *Sex Roles*, vol. 86, no. 1–2, pp. 89–105, Jan. 2022, doi: 10.1007/S11199-021-01254-1/TABLES/5.
4. D. M. Doumas and A. Midgett, “The association between witnessing cyberbullying and depressive symptoms and social anxiety among elementary school students”, *Psychol Sch*, vol. 58, no. 3, pp. 622–637, Mar. 2021, doi: 10.1002/PITS.22467.
5. D. Uludasdemir and S. Kucuk, “Cyber Bullying Experiences of Adolescents and Parental Awareness: Turkish Example”, *J Pediatr Nurs*, vol. 44, pp. e84–e90, Jan. 2019, doi: 10.1016/J.PEDN.2018.11.006.
6. E. Duman and N. Yilmaz, “Investigation of the Effect of Cyber Bullying Awareness Training for Parents 1”, vol. 84, no. 3, pp. 406–421, 2021, doi: 10.38089/ekuad.2021.84.
7. S. X. Zhang and E. Van Burg, “Advancing entrepreneurship as a design science: developing additional design principles for effectuation”, *Small Business Economics*, vol. 55, no. 3, 2020, doi: 10.1007/s11187-019-00217-x.
8. B. Xiao, N. Parent, T. Bond, J. Sam, and J. Shapka, “Developmental Trajectories of Cyber-Aggression among Early Adolescents in Canada: The Impact of Aggression, Gender, and Time Spent Online”, *Int J Environ Res Public Health*, vol. 21, no. 4, Apr. 2024, doi: 10.3390/ijerph21040429.
9. R. Lozano-Blasco, A. Cortés-Pascual, and M. P. Latorre-Martínez, “Being a cybervictim and a cyberbully – The duality of cyberbullying: A meta-analysis”, *Comput Human Behav*, vol. 111, Oct. 2020, doi: 10.1016/j.chb.2020.106444.
10. C. Nixon, “Current perspectives: the impact of cyberbullying on adolescent health”, *Adolesc Health Med Ther*, p. 143, Aug. 2014, doi: 10.2147/AHMT.S36456.
11. R. M. Kowalski and S. P. Limber, “Psychological, physical, and academic correlates of cyberbullying and traditional bullying”, *Journal of Adolescent Health*, vol. 53, no. 1 SUPPL, 2013.
12. Q. Chen, K. L. Chan, S. Guo, M. Chen, C. K. ming Lo, and P. Ip, “Effectiveness of Digital Health Interventions in Reducing Bullying and Cyberbullying: A Meta-Analysis”, *Trauma Violence Abuse*, vol. 24, no. 3, pp. 1986–2002, Jul. 2023, doi: 10.1177/15248380221082090.

13. H. Gaffney, D. P. Farrington, and M. M. Ttofi, "Examining the Effectiveness of School-Bullying Intervention Programs Globally: a Meta-analysis", *International Journal of Bullying Prevention*, vol. 1, no. 1, pp. 14–31, 2019, doi: 10.1007/s42380-019-0007-4.
14. R. Khan, "Standardized Architecture for Conversational Agents a.k.a. ChatBots", *International Journal of Computer Trends and Technology*, vol. 50, no. 2, pp. 114–121, Aug. 2017, doi: 10.14445/22312803/IJCTT-V50P120.
15. N. Lessio and A. Morris, "Toward Design Archetypes for Conversational Agent Personality", *Conf Proc IEEE Int Conf Syst Man Cybern*, vol. 2020-October, pp. 3221–3228, Oct. 2020, doi: 10.1109/SMC42975.2020.9283254.
16. A. Følstad, M. Skjuve, and P. B. Brandtzaeg, "Different chatbots for different purposes: Towards a typology of chatbots to understand interaction design", *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 11551 LNCS, pp. 145–156, 2019, doi: 10.1007/978-3-030-17705-8_13.
17. S. Hobert, A. Følstad, and E. L. C. Law, "Chatbots for active learning: A case of phishing email identification", *International Journal of Human Computer Studies*, vol. 179, Nov. 2023, doi: 10.1016/J.IJHCS.2023.103108.
18. F. D. Davis, "Perceived usefulness, perceived ease of use, and user acceptance of information technology", *MIS Quarterly*, vol. 13, no. 3, pp. 318–339, 1989, doi: 10.2307/249008.
19. M. Turner, B. Kitchenham, P. Brereton, S. Charters, and D. Budgen, "Does the technology acceptance model predict actual use? A systematic literature review", *Inf Softw Technol*, vol. 52, no. 5, pp. 463–479, May 2010, doi: 10.1016/j.infsof.2009.11.005.
20. P. Kowalcuk, "Consumer acceptance of smart speakers: a mixed methods approach", *Journal of Research in Interactive Marketing*, vol. 12, no. 4, pp. 418–431, Oct. 2018, doi: 10.1108/JRIM-01-2018-0022.
21. V. Venkatesh, M. G. Morris, G. B. Davis, and F. D. Davis, "User acceptance of information technology: Toward a unified view", *MIS Quarterly*, vol. 27, no. 3, pp. 425–478, 2003, doi: 10.2307/30036540.
22. S. Alharbi, "Trust and acceptance of cloud computing: A revised UTAUT model", *Computational Science and Computational*, 2014, Accessed: Aug. 22, 2016. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6822317
23. Y. K. Dwivedi, N. P. Rana, A. Jeyaraj, M. Clement, and M. D. Williams, "Re-examining the Unified Theory of Acceptance and Use of Technology (UTAUT): Towards a Revised Theoretical Model", *Information Systems Frontiers*, vol. 21, no. 3, pp. 719–734, Jun. 2019, doi: 10.1007/S10796-017-9774-Y.
24. L. Labadze, M. Grigolia, and L. Machaidze, "Role of AI chatbots in education: systematic literature review", *International Journal of Educational Technology in Higher Education*, vol. 20, no. 1, p. 56, Oct. 2023, doi: 10.1186/s41239-023-00426-1.
25. L. Lengel, Robert, H. and Daft, Richard, "The Selection of Communication Media as an Executive Skill", *Academy of Management Perspective*, vol. 2, no. 3, pp. 225–232, 1988.
26. V. C. Sheer, "Media Richness Theory", *The International Encyclopedia of Media Psychology*, pp. 1–14, Sep. 2020, doi: 10.1002/9781119011071.IEMP0118.
27. S. Gabrielli, S. Rizzi, S. Carbone, and V. Donisi, "A chatbot-based coaching intervention for adolescents to promote life skills: Pilot study", *JMIR Hum Factors*, vol. 7, no. 1, Mar. 2020, doi: 10.2196/16762.
28. S. Gabrielli, S. Rizzi, S. Carbone, and V. Donisi, "A Chatbot-Based Coaching Intervention for Adolescents to Promote Life Skills: Pilot Study", *JMIR Hum Factors*, vol. 7, no. 1, p. e16762, Feb. 2020, doi: 10.2196/16762.

29. S. Herkama and C. Salmivalli, “KiVa antibullying program”, *Reducing Cyberbullying in Schools: International Evidence-Based Best Practices*, pp. 125–134, Jan. 2017, doi: 10.1016/B978-0-12-811423-0.00009-2.
30. C. F. Garandeau, E. Poskiparta, and C. Salmivalli, “Tackling acute cases of school bullying in the KiVa anti-bullying program: A comparison of two approaches”, *J Abnorm Child Psychol*, vol. 42, no. 6, pp. 981–991, 2014, doi: 10.1007/S10802-014-9861-1.
31. Y. You, C. H. Tsai, Y. Li, F. Ma, C. Heron, and X. Gui, “Beyond Self-diagnosis: How a Chatbot-based Symptom Checker Should Respond”, *ACM Transactions on Computer-Human Interaction*, vol. 30, no. 4, Sep. 2023, doi: 10.1145/3589959.
32. D. B. Jackson and M. G. Vaughn, “The bully-victim overlap and nutrition among school-aged youth in North America and Europe”, *Child Youth Serv Rev*, vol. 90, pp. 158–165, Jul. 2018, doi: 10.1016/j.childyouth.2018.05.027.
33. A. Cosma *et al.*, “Bullying victimization: time trends and the overlap between traditional and cyberbullying across countries in Europe and North America”, *Int J Public Health*, vol. 65, no. 1, pp. 75–85, Jan. 2020, doi: 10.1007/s00038-019-01320-2.
34. J. W. Creswell and V. L. P. Clark, *Designing and conducting mixed methods research*. Sage publications, 2017.
35. A. Yoachim, E. T. Welty, and N. Jenisch, “Public Interest Design, Pragmatism, and Potentials in a Postdiluvian City”, *Journal of Architectural Education*, vol. 74, no. 1, pp. 110–118, 2020, doi: 10.1080/10464883.2020.1693834.
36. D. Chiffi and M. Andreoletti, “Abduction in Prognostic Reasoning”, *Handbook of Abductive Cognition*, pp. 1–23, 2022, doi: 10.1007/978-3-030-68436-5_11-1/COVER.
37. D. Walton, *Abductive reasoning*. University of Alabama Press, 2014.
38. H. C. (Oliver) Chan and D. S. W. Wong, “Coping with cyberbullying victimization: An exploratory study of Chinese adolescents in Hong Kong”, *Int J Law Crime Justice*, vol. 50, pp. 71–82, Sep. 2017, doi: 10.1016/j.ijlcj.2017.04.003.
39. A. Hevner *et al.*, “Design science in information systems research”, *MIS quarterly*, vol. 28, no. 1, pp. 75–105, 2004, doi: 10.2307/25148625.
40. R.; Baskerville *et al.*, “Design science research contributions Design Science Research Contributions: Finding a Balance between Artifact and Theory”, *J Assoc Inf Syst*, vol. 19, no. 5, pp. 358–376, 2018, doi: 10.17705/1jais.00495.
41. A. R. Hevner, S. T. March, J. Park, and S. Ram, “Design science in information systems research”, *MIS Q*, vol. 28, no. 1, pp. 75–105, 2004, doi: 10.2307/25148625.
42. M. L. Ranney *et al.*, “Cyberbullying Prevention for Adolescents: Iterative Qualitative Methods for Mobile Intervention Design”, *JMIR Form Res*, vol. 5, no. 8, p. e25900, Aug. 2021, doi: 10.2196/25900.
43. B. Zhang, J. Luo, and J. Li, “Moving beyond Likert and Traditional Forced-Choice Scales: A Comprehensive Investigation of the Graded Forced-Choice Format”, *Multivariate Behav Res*, vol. 59, no. 3, pp. 434–460, 2024, doi: 10.1080/00273171.2023.2235682.
44. B. Çetin, E. Yaman, and A. Peker, “Cyber victim and bullying scale: A study of validity and reliability”, vol. 57, no. 4, pp. 2261–2271, Dec. 2011, doi: 10.1016/J.COMPEDU.2011.06.014.
45. J. Rose and C. W. Johnson, “Contextualizing reliability and validity in qualitative research: toward more rigorous and trustworthy qualitative social science in leisure research”, *J Leis Res*, vol. 51, no. 4, pp. 432–451, Aug. 2020, doi: 10.1080/00222216.2020.1722042.
46. N. Humaidi and V. Balakrishnan, “Exploratory Factor Analysis of User? s Compliance Behaviour towards Health Information System? s Security”, *Journal of Health & Medical*,

- 2013, Accessed: Aug. 22, 2016. [Online]. Available: <http://www.omicsonline.org/exploratory-factor-analysis-of-user-compliance-behaviour-towards-health-2157-7420.1000123.php?aid=11780>
- 47. J. M. Conway and A. I. Huffcutt, “A Review and Evaluation of Exploratory Factor Analysis Practices in Organizational Research”, Apr. 2003. doi: 10.1177/1094428103251541.
 - 48. M. Schrepp, “On the Usage of Cronbach’s Alpha to Measure Reliability of UX Scales”, *J Usability Stud*, vol. 15, pp. 247–258, 2020.
 - 49. E. Adamopoulou and L. Moussiades, “Chatbots: History, technology, and applications”, *Machine Learning with Applications*, vol. 2, p. 100006, Dec. 2020, doi: 10.1016/J.MLWA.2020.100006.
 - 50. A. Hevner and S. Chatterjee, “Design Science Research in Information Systems”, 2010, pp. 9–22. doi: 10.1007/978-1-4419-5653-8_2.

Using AI to boost scoping reviews. Exploring AI deployment in obstetrics and gynaecology as an exemplar

Isabella M. Venter^{1[0000-0001-6554-8024]}, Rénette J. Blignaut^{1[0000-0002-3150-6624]} and Karen V. Renaud^{2,3 [0000-0002-7187-6531]}

¹ University of the Western Cape, Cape Town, South Africa

² University of Strathclyde, Glasgow, Scotland

³ University of South Africa, Pretoria, South Africa

rblignaut@uwc.ac.za; karen.renaud@strath.ac.uk

Abstract. This paper explores how artificial intelligence can be harnessed to boost and supercharge scoping reviews. To demonstrate its power (comprehensiveness), its efficiency (time effectiveness) and affordability (cost effectiveness), we explore the use of AI in obstetrics and gynaecology as a vehicle to test our new hybrid review process. We carried out: (1) a traditional scoping review, and (2) a traditional scoping review boosted by artificial intelligence tools. We provide an overview of the extensive and varied literature related to our topic of interest: the use of AI in obstetrics and gynaecology. We compare and contrast the outcomes of the different options. We conclude that the use of AI tools can enrich and extend the scope of scoping reviews (using a so-called “Hybrid Review”), but only if the prompts are carefully and thoughtfully crafted.

Keywords: artificial intelligence, large language models, chatbots, obstetrics, gynaecology, diagnosis, prognosis, ethics.

1 Introduction

Natural language processing (NLP)—a field in artificial intelligence (AI)—is being used by Large Language Models (LLMs) across numerous fields, including healthcare. LLMs are trained on massive datasets and use billions of parameters to perform tasks such as answering questions, translating text, and completing sentences [1]. Models such as the Generative Pre-trained Transformer (GPT) [2] and Gemini [3] have demonstrated the ability to produce coherent and contextually appropriate text. Since the introduction of the chatbot ChatGPT by OpenAI on 21 November 2022, many similar bots have been developed. Poe.com, for example, is a website that lists several of these bots: official bots, reasoning bots, search bots, image generating bots, text analysis bots, etc. These AI-powered bots generate responses dynamically, often using large language models or external databases for context.

However, these AI tools’ use in research and practice is constrained by significant challenges, such as their tendency to generate inaccurate or misleading information, often referred to as “hallucination” [4]. Furthermore, LLM’s are time constrained—i.e.,

it is limited by the training data which is at best a few months old [5] and thus the information they provide is often not current.

Retrieval-Augmented Generation (RAG) is a method that enhances the output of a large language model by integrating information from an authoritative external knowledge base, beyond the model's original training data, before generating a response [6]. RAG builds on the powerful capabilities of LLMs by enabling them to access domain-specific information without requiring model retraining. This makes RAG a cost-effective solution for improving the relevance, accuracy, and applicability of LLM outputs across a variety of contexts (see Fig. 1). For example, in healthcare, RAG can be utilised to handle requests for relevant patient records or research papers and generate personalised treatment plans. For example, it will be able to analyse a patient's medical history and combine this with recent research insights to recommend tailored treatment solutions just prior to the initiation of care [5].

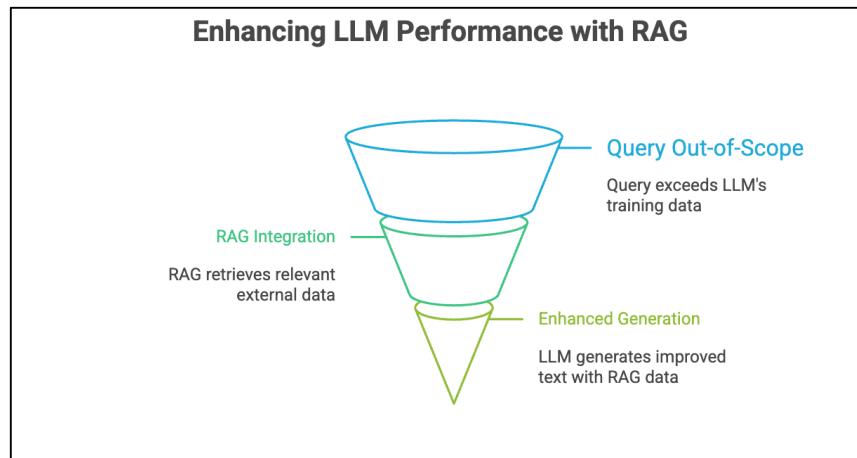


Fig. 1: Example of how RAG is used to add recent or more specific information (generated by Napkin.ai)

Although AI has been adopted in academia and in education, many faculty members, especially in the health sciences, are still unsure of its benefits or how to advise its ethical deployment. As reported in a recent study “*There was broad consensus among faculty members that students should receive training on how to use AI chatbots*” [7] (p. 5). Haleem *et al.* [8] argue that the use of AI is going to change all domains of medicine. According to Weidener and Fischer [9] although many health science students engage with AI technology, particularly AI chat applications, only a small percentage have undergone formal AI training. This indicates a significant gap in current health science curricula. It underscores the need to update health science education to include AI and AI ethics, ensuring that future healthcare professionals are well-prepared to address the challenges and opportunities that AI presents in healthcare [9].

Scoping reviews are incredibly powerful in terms of giving researchers a snapshot of what researchers have found related to any topic [12]. Traditional scoping reviews have strong power in terms of delivering a comprehensive set of papers. It is also affordable in terms of the tool being free to anyone who wants to use it. However, it is extremely time consuming. AI-powered reviews, on the other hand, are extremely efficient—identifying papers in a matter of minutes. However, to be able to harness their full power, researchers must pay for the premier version. As such, these tools are not universally affordable to many researchers.

Fig. 2 shows how the different review tools perform on three measures: power (comprehensiveness), affordability (cost effectiveness) and efficiency (time effectiveness). Traditional scoping reviews are affordable, given that most universities provide access to researchers. If carried out rigorously, it has maximum power to find all relevant publications. It is, however, extremely time consuming, so does not exhibit efficiency. AI-Powered reviews are very efficient, delivering their outcomes in a matter of minutes. To get the full benefit of their power, a paid subscription is usually required, which reduces their affordability. They do not have a full measure of power, given that they do not reveal the same comprehensive coverage as traditional reviews. A hybrid review, on the other hand, can deliver the best of both worlds.

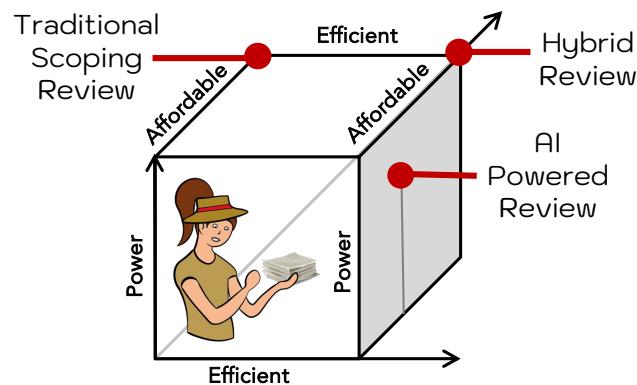


Fig. 2: Characteristics of scoping review tools (drawn by the authors)

This study examined how artificial intelligence can be used to discover its deployment in healthcare with specific reference to obstetrics and gynaecology as proof of concept. The research question being posed is: *How can AI help streamline the scoping review process of identifying current AI applications in all domains?*

This question can be translated into two sub-questions:

RQ1. *To what extent can hybrid tools (traditional searches augmented with AI) boost scoping reviews, in terms of power, efficiency and affordability?*

RQ2. *How has AI been used in obstetrics and gynaecology?*

A scoping review and a review using AI tools were used to address the first research question. The results were compared, and it was found that AI tools simplify the search and identification of appropriate papers considerably. The only limitation being the AI tools' affordability. Furthermore, using AI to do the content analysis it was found that AI is being used mostly for diagnostic imaging and predictive analytics within the field of obstetrics and gynaecology. The main contribution of this paper is to suggest a hybrid review process, harnessing both traditional and AI-powered searches. We tested our new process by searching for papers in one specific branch of medicine, but it will apply equally to other searches.

2 Materials & Methods

Pragmatism, as a research approach, focuses on understanding the world and addressing its challenges by combining both subjective and objective methods. Pragmatist researchers adopt a flexible and practical approach to finding effective solutions for the phenomena being studied and acknowledge the value of incorporating diverse perspectives and worldviews to tackle research problems [10].

In higher education, especially in the context of AI, pragmatism offers a framework for educators and institutions to navigate technology-enhanced learning environments. This perspective aligns with John Dewey's educational theories, which stress that knowledge should be rooted in real-world experiences and that education should prepare individuals for practical life [11].

2.1 Methodology

A pragmatic approach—which emphasises the practical implications of research findings and integrating various methods to address the research question—was used for this research. The methods employed were a scoping review using PRISMA and a review using AI tools—SciSpace, Elite Beta and Consensus. These AI tools were randomly chosen from AI tools that offer free options. Finally, the results were compared (see **Fig. 3**) and content analysis was used to address the research questions.

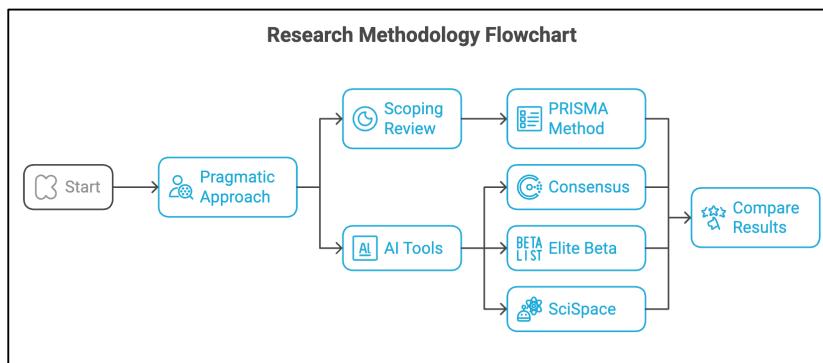


Fig. 3 Research approach (defined by the researchers and generated by Napkin.ai)

2.2 Methods

The scoping review as method, is suitable since “...scoping reviews do not aim to produce a critically appraised and synthesised result/answer to a particular question, ...it rather aims to provide an overview or map of the evidence.” [12] (p. 3).

Scoping review using PRISMA. The Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) was primarily intended for systematic reviews of research evaluating the effects of health interventions but has since been adopted by many other fields of study (see Fig. 4).

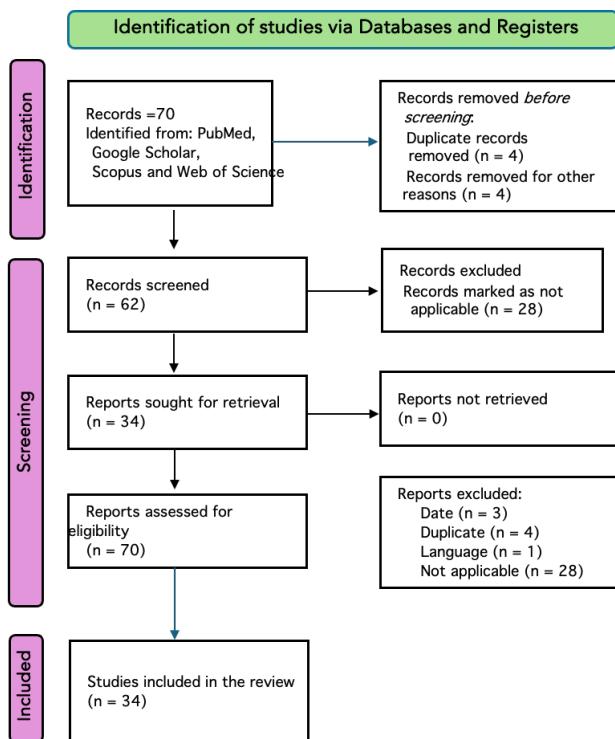


Fig. 4: PRISMA flowchart of articles identified, screened and included in the study

As can be seen from the figure, it has three main steps for reviewing literature: identification of articles, screening of the articles and deciding which articles to include in the review.

Using the search word/phrases: AI and obstetrics, AI and gynaecology, AI and obstetrics and gynaecology, AI and obstetrics and gynaecology and education, 70 papers that could be downloaded in full were identified by the researchers using scholarly databases. Twenty-five articles were found using PubMed, 18 using Google Scholar, 18

using Scopus and 9 using Web of Science. All of these were published in the period 2024–2025.

No automation tools—such as Rayyan—were used to generate the PRISMA flowchart, all the literature were evaluated by two researchers. Of the 70 articles only 34 were retained for review. Four articles were removed since they were duplicates, another 4 were removed because of incorrect publication dates and one being written in German. Twenty-eight articles were removed because they dealt with medical education but not specifically with the training of obstetrics and gynaecology medical students or registrars. This search and identification of relevant papers was carried out over several days and took approximately 10 hours altogether.

Review using AI tools. It was decided to implement some freely available AI tools such as SciSpace, Elicit Beta, and Consensus to identify relevant articles. More AI tools are available such as: Semantic Scholar, Research Rabbit, Connected, and LitSence. etc. these were however not used in this investigation.

SciSpace, was interrogated—on the 20th of April 2025—to identify journal articles for the first research question, namely “*How is AI being used in obstetrics and gynaecology in terms of diagnosis, its challenges and benefits?*” SciSpace has three levels of searches, namely Standard, High Quality or Deep Review. Deep Review was chosen and 1050 papers were identified of which it suggested 243 were relevant to the research question. It furthermore wrote a short summary of the “top” 20 papers—it seems as if it refers to the most recent papers identified—in terms of practical applications in clinical settings, theoretical advancements and research and development: It could also be instructed to write a summary of the top 10 or 5 papers. When the same request was made on the 30th June 2025, only the standard review could be accessed without payment. The search identified 100 relevant journal papers with a summary of the first 5 papers without payment. For some articles the PDF could be requested, for others it could be downloaded directly. This search was used for the comparison.

Elicit Beta allowed the user to use some filters—for dates and quality of publications—to indicate what the search should include. When it was instructed (on the 20th of April 2025) to search for papers using the first research question it rephrased the research question to: “*What are the primary ethical, technical, and implementation challenges of integrating AI technologies in obstetric and gynaecological diagnostic practices?*” The free version allowed the consideration of 50 papers. However, according to the tool not all these papers’ full texts were checked—only the abstract was checked for some of these identified papers. In its report Elicit Beta indicated that it searched across over 126 million academic papers from the Semantic Scholar corpus and retrieved the 50 papers most relevant to the query. On the 2nd of June 2025 when the research was repeated for the question “*What are the current applications of artificial intelligence in obstetrics and gynaecology for diagnosis and treatment?*” it identified 50 papers which according to it, is most relevant to the research question. The second search was used for the comparison.

When the same question was posed to Consensus, it provided 10 papers related to the research question as well as a summary of the content of the papers. Filters could be used and more papers—than the ten summarised—could be loaded. Information

about the papers such as the study population, study count, methods, outcomes and results could be retrieved. Consensus also allows for the download of a CSV file with a summary of the papers. The summary of all the information about the 10 papers was useful and was used for the comparison (see Fig. 5).

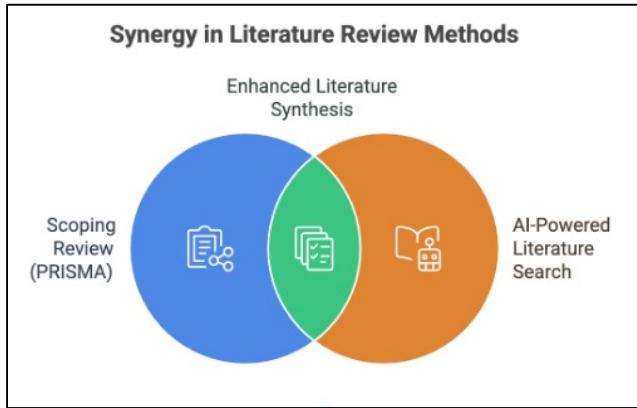


Fig. 5: Comparing the different methods (generated by Napkin.AI)

Comparison. The process using the AI tools was efficient —much more so than the traditional scoping review process. For the articles found, all the tools—SciScape, Elite Beta and Consensus—provided links which could be used to access the articles. That is, to confirm that the AI tools had not “*hallucinated*” and that these references and articles actually existed.

3 Results

It was decided to see which articles identified with the PRISMA method were also identified by the AI tools. As can be seen from Table 1, nine of the 34 articles identified using the PRISMA method, was also mentioned by one or more of the AI tools. Six was mentioned by Consensus, three by Elite Beta and 5 by SciSpace. Four articles were mentioned by three of the approaches. Only one article was not identified during the PRISMA search but was mentioned by all three of the AI tools.

Table 1: Comparing the PRISMA results with the results when using AI tools

	<i>Title of the Article</i>	<i>DOI</i>	<i>PRISMA</i>	<i>Con-sensus</i>	<i>Elite Beta</i>	<i>Sci-Space</i>
1	Elbiss et al. Artificial intelligence in gynecologic and obstetric emergencies (2025)	10.1186/s12245-025-00820-8	x	x		
2	Saida, et al. Artificial Intelligence in Obstetric and Gynecological MR Imaging (2024)	10.2463/mrms.rev.2024-0077	x	x	x	
3	Patel et al. Artificial Intelligence in Obstetrics and Gynecology: Transforming Care and Outcomes (2024)	10.7759/cureus.64725	x	x		x

	<i>Title of the Article</i>	<i>DOI</i>	<i>PRISMA</i>	<i>Con-sensus</i>	<i>Elite-Beta</i>	<i>Sci-Space</i>
4	Sone et al. Clinical Prospects for Artificial Intelligence in Obstetrics and Gynecology (2024)	10.31662/jma.j.2024-0197	x	x		
5	Changhez et al. Evaluating the Efficacy and Accuracy of AI Assisted Diagnostic Techniques in Endometrial Carcinoma: A Systematic Review (2024)	10.7759/cureus.60973		x		x
6	Moro et al. Role of artificial intelligence applied to ultrasound in gynecology oncology: A systematic review (2024)	10.1002/ijc.35092		x	x	x
7	Brandão et al. Revolutionizing Women's Health: A Comprehensive Review of Artificial Intelligence Advancements in Gynecology (2024)	10.3390/jcm13041061		x	x	
8	Gumilar et al. The promise and challenges of Artificial Intelligence-Large Language Models (AI-LLMs) in obstetric and gynecology (2024)	10.20473/mo.g.v32i22024.128-135		x		x
9	Mahmoudiandehkordi et al. Transforming Gynecology with Artificial Intelligence: Advances in Clinical Practice (2024)	10.54756/ijstar.2024.25		x		x
10	Chaurasia et al. Use of artificial intelligence in obstetric and gynaecological diagnostics: a protocol for a systematic review and meta-analysis (2024)	10.1136/bmjopen-2023-082287		x	x	x

Content analysis. Using ChatPDF, these ten papers were summarised. The summary of all ten papers where then used to prompt ChatGPT, Grok 3 and Deepseek, to extract the broad themes of the research. The researchers read through the summarised ten papers and identified specific themes.

As can be seen from **Table 2**, two cross-cutting themes were identified by all three AI tools and the researchers. These were: AI in diagnostic imaging and predictive analysis.

Table 2: Broad themes of the research

<i>Theme</i>	<i>Grok 3</i>	<i>Deepseek</i>	<i>ChatGPT</i>	<i>Researchers</i>
AI in diagnostic imaging	x	x	x	x
Predictive analytics	x	x	x	x
Cancer detection & staging	x	x		
Personalised medicine	x		x	x
Ethical and practical considerations	x		x	x
Reproductive medicine		x		
Emergency care		x		
Treatment planning			x	x

<i>Theme</i>	<i>Grok 3</i>	<i>Deepseek</i>	<i>ChatGPT</i>	<i>Researchers</i>
Challenges such as data quality, lack of standardisation, interpretability issues, and limited clinical validation	x	x		
Human oversight			x	

Personalised medicine and ethical considerations were mentioned by two of these tools and the researchers. Treatment planning and challenges such as data quality, lack of standardisation, interpretability issues, and limited clinical validation was mentioned by ChatGPT and the researchers. Other themes mentioned were AI's use for reproductive medicine, emergency care and the need for human oversight.

4 Discussion

This research considered how AI (using AI tools) can help streamline scoping reviews to identify current AI applications in all domains (see Fig. 6).

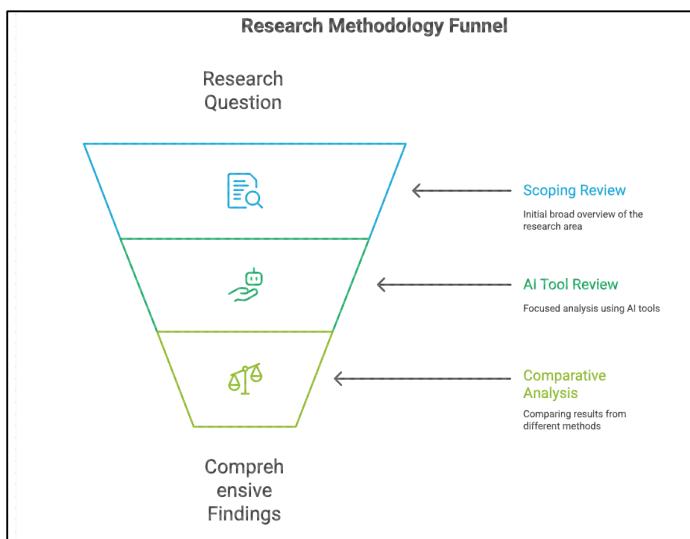


Fig. 6: Results using the research methodology funnel

4.1 Hybrid tools

The scoping review, using the PRISMA method, took approximately 10 hours and produced 34 articles that could be analysed. Whereas the review, using AI Tools, produced 160 papers in a matter of minutes providing—in most cases—a summary of the first 10 papers.

When all these papers were compared—those identified by both the traditional PRISMA method and using AI tools—it was found that nine papers were identified by at least one of the AI tools and PRISMA. To answer the first research question:

RQ1. *To what extent can hybrid tools (traditional searches augmented with AI) boost scoping reviews, in terms of power, efficiency and affordability?*

It became clear that using a hybrid approach can boost scoping reviews in terms of power—many relevant articles can be found in a short period, efficient in that ten of the papers were identified by most and affordable in that none of the tools used required a subscription.

Our conclusion is that AI tools can enrich and extend the scope of scoping reviews, but only if the prompts are carefully and thoughtfully crafted.

4.2 AI use in obstetrics and gynaecology

Analysing the ten papers that were identified, the second research question was addressed:

RQ2. *How has AI been used in obstetrics and gynaecology?*

It was found that AI could indeed assist with personalised medicine, ethical considerations, treatment planning but that there are challenges such as data quality, lack of standardisation, interpretability issues, and limited clinical validation. AI were mostly used in reproductive medicine, for cancer treatment and in emergency care. However, all of these require human oversight.

The question that may be asked when using AI for scoping reviews is: Are these identified articles representative of the research field globally? Furthermore, in terms of prompt engineering researchers need to be aware that when using an AI tool it does not remove bias from underlying training data and also that slight changes in prompts can yield different outputs, therefore it is important to test multiple prompt versions for reliability.

4.3 Future work

We suggest the following avenues for future research:

1. AI services usually only provide the first 10 results free of charge. The consequent lack of access to paid versions has a negative impact on comprehensiveness. Investigations into the impacts of different AI-tool pricing strategies would service to refine our hybrid review process.
2. In fleshing out the hybrid review approach, it would be good to explore the inclusion of an “expert in the loop” to help to decide on which tools to use and which papers to include in the final corpus. This would have an impact on power but is likely to impact time effectiveness.

Acknowledgements

The first author sadly passed away a week before the paper's deadline. Since she had spearheaded this work, and had worked on the paper the day before she passed, her family wanted us to complete and submit the paper to honour her. The second and third authors express their appreciation for many years of valued collaborations with Isabella, for her friendship and her tremendous sense of fun, which made all interactions such a great pleasure. We are pleased to be able to present this paper as her final contribution to the research literature.

List of Papers from Review

1. Elbiss, H. M., Fikri M. A-Z.: Artificial intelligence in gynecologic and obstetric emergencies. International Journal of Emergency Medicine 18(1), Article 20 (2025). 10.1186/s12245-025-00820-8
2. Saida, T., Wenchao G., Sodai H., Toshitaka I., Masafumi S., Taishi A., Yuta N., Ayumi S., Toyomi S., Takahito N.: Artificial Intelligence in Obstetric and Gynecological MR Imaging. Magnetic Resonance in Medical Sciences 24(3), 354-365 (2025). 10.2463/mrms.rev.2024-0077
3. Patel, D. J., Chaudhari, K., Acharya, N., Shrivastava, D., Muneeba, S.: Artificial intelligence in obstetrics and gynecology: transforming care and outcomes. Cureus 16(7), e64725 (2024). 10.7759/cureus.64725
4. Sone, K., Taguchi, A., Miyamoto, Y., Uchino-Mori, M., Iriyama, T., Hirota, Y., Osuga, Y.: Clinical Prospects for Artificial Intelligence in Obstetrics and Gynecology. JMA Journal 8(1), 113-120 (2025) 10.31662/jmaj.2024-0197
5. Changhez, J., James, S., Jamala, F., Khan, S., Khan, M. Z., Gul, S., Zainab, I.: Evaluating the efficacy and accuracy of AI-assisted diagnostic techniques in endometrial carcinoma: A systematic review. Cureus 16(5), e60973 (2024). 10.7759/cureus.60973
6. Moro, F., Ciancia, M., Zace, D., Vagni, M., Tran, H. E., Giudice, M. T., Zoccoli, Mascilini, S. G. F., Ciccarone, F., Boldrini, L., D'Antonio, F., Scambia, G., Testa, A. C.: Role of artificial intelligence applied to ultrasound in gynecology oncology: A systematic review. International journal of cancer 155(10), 1832-1845 (2024). 10.1002/ijc.35092
7. Brandão, M., Mendes, F., Martins, M., Cardoso, P., Macedo, G., Mascarenhas, T., Saraiva, M. M.: Revolutionizing women's health: A comprehensive review of artificial intelligence advancements in gynecology. Journal of Clinical Medicine 13(4), 1061 (2024). 10.3390/jcm13041061
8. Gumilar, K. E., Tan, M.: The promise and challenges of Artificial Intelligence-Large Language Models (AI-LLMs) in obstetrics and gynecology. Journal Of Obstetrics & Gynecology Science 136(2), 128-135 (2024). 10.20473/mog.v32i22024.128-135
9. Mahmoudiandehkordi, S., Yeganegi, M. Transforming Gynecology with Artificial Intelligence: Advances in Clinical Practice. International Journal of Scientific and Applied Research (IJSAR), 4(9), 104-112 (2024). 10.54756/ijsar.2024.25

10. Chaurasia, A., Curry, G., Zhao, Y., Dawoodbhoy, F., Green, J., Vaninetti, M., Shah, N., Greer, O.: Use of artificial intelligence in obstetric and gynaecological diagnostics: a protocol for a systematic review and meta-analysis. *BMJ Open* 14(5), e082287 (2024). 10.1136/bmjopen-2023-082287

References

- [1] Floridi, L., Chiriatti, M.: GPT-3: Its Nature, Scope, Limits, and Consequences. *Minds & Machines*, 30, 681–694. (2020). 10.1007/s11023-020-09548-1
- [2] OpenAI. (2023). GPT-4 Technical Report. Retrieved from: <https://arxiv.org/abs/2303.08774>.
- [3] Google. (2023, December 6). Gemini. Retrieved from: Gemini.google.com: <https://gemini.google.com>
- [4] Huang, L., W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, Q. Chen, W. Peng, X. Feng, B. Qin, T. Liu.: A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems* 43(2), 1-55. (2025). 10.1145/3703155
- [5] Amugongo, L. M., Mascheroni, P., Brooks, S., Doering, S., Seidel, J.: Retrieval augmented generation for large language models in healthcare: A systematic review. *PLOS Digital Health* 4(6), e0000877 (2025). 10.1371/journal.pdig.0000877
- [6] Fan, W., Ding, Y., Ning, L., Wang, S., Li, H., Yin, D., T-S Chua, Li, Q. A survey on rag meeting LLMs: Towards retrieval-augmented large language models. In: Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pp. 6491-6501. Barcelona, Spain: ACM. (2024). 10.1145/3637528.3671470
- [7] Saleh, Z. T., R. A. Elshatarat, M. Alharbi, B. N. Alhumaidi, M. S. Al-Za'areer, R. A. Jarrad, T. F. Al Niarat, W. T. Almagharbeh, K. M. Al-Sayaghi, D. El Sayed Fadila. Exploring faculty perceptions and concerns regarding artificial intelligence Chatbots in nursing education: potential benefits and limitations. *BMC Nursing*, 24(440), 1-13 (2025). 10.1186/s12912-025-03082-0
- [8] Haleem, A., Javaid, M. & Khan, I.H. Current status and applications of Artificial Intelligence (AI) in medical field: An overview. *Current Medicine Research and Practice* 9(6), 231-237 (2019). 10.1016/j.cmrp.2019.11.005
- [9] Weidener, L., & Fischer, M.: Artificial intelligence in medicine: cross-sectional study among medical students on application, education, and ethical aspects. *JMIR medical education*, 10(1), e51247. (2024). 10.2196/51247
- [10] Elgeddawy, M., & Abouraia, M.: Pragmatism as a research paradigm. European Conference on Research Methodology for Business and Management Studies (pp. 71-74). Academic Conferences International Limited. (2024)
- [11] Lin, F., Chiew Hwa, P., & Ku Wing, C.: The Influence of Dewey's Pragmatism Educational Thought on Chinese General Education: A Historical Summary. *Malaysian Journal of Social Sciences and Humanities*, 8(12), p. e002589 (2023). 10.47405/mjssh.v8i12.2589
- [12] Munn, Z. P. Systematic review or scoping review? Guidance for authors when choosing between a systematic or scoping review approach. *BMC Medical Research Methodology*, 18(143), 1-7. (2018). 10.1186/s12874-018-0611-x

A Cross-Cultural Assessment of Human Ability to Detect LLM-Generated Fake News about South Africa

Tim Schlippe¹[0000-0002-9462-8610], Matthias Wölfel²[0000-0003-1601-5146],
Koena Ronny Mabokela³[0000-0002-8058-969X]

¹ IU International University of Applied Sciences, Germany

² Karlsruhe University of Applied Sciences, Germany

³ Applied Information Systems, University of Johannesburg, South Africa
tim.schlippe@iu.org

Abstract. This study investigates how cultural proximity affects the ability to detect AI-generated *fake* news by comparing South African participants with those from other nationalities. As large language models increasingly enable the creation of sophisticated *fake* news, understanding human detection capabilities becomes crucial, particularly across different cultural contexts. We conducted a survey where 89 participants (56 South Africans, 33 from other nationalities) evaluated 10 *true* South African news articles and 10 AI-generated *fake* versions. Results reveal an asymmetric pattern: South Africans demonstrated superior performance in detecting *true* news about their country (40% deviation from ideal rating) compared to other participants (52%), but performed worse at identifying *fake* news (62% vs. 55%). This difference may reflect South Africans' higher overall trust in news sources. Our analysis further shows that South Africans relied more on content knowledge and contextual understanding when judging credibility, while participants from other countries emphasised formal linguistic features such as grammar and structure. Overall, the deviation from ideal rating was similar between groups (51% vs. 53%), suggesting that cultural familiarity appears to aid verification of authentic information but may also introduce bias when evaluating fabricated content. These insights contribute to understanding cross-cultural dimensions of misinformation detection and inform strategies for combating AI-generated *fake* news in increasingly globalised information ecosystems where content crosses cultural and geographical boundaries.

Keywords: Fake News · Disinformation · Natural Language Processing · South Africa · Large Language Models · LLMs.

1 Introduction

Large language models (LLMs) have dramatically lowered barriers to producing convincing *fake* news [?,?]. Unlike traditional misinformation with obvious

flaws, AI-generated content is coherent, sophisticated, and increasingly indistinguishable from *true* news [?,?]. This technological shift creates an urgent need to understand how humans detect—or fail to detect—AI-generated *fake* news.

This challenge intensifies when readers lack cultural or factual familiarity with the subject matter, as they struggle to identify inaccuracies that would normally serve as red flags [?,?]. This cultural dimension of *fake* news detection remains underexplored, particularly for AI-generated content about regions with less global media attention, such as South Africa [?,?].

This study investigates the influence of cultural proximity on the identification of AI-generated *fake* news by comparing South African participants with contextual knowledge against those less acquainted with South African issues. This reveals how background knowledge influences misinformation detection—crucial as AI content proliferates globally [?,?]. Our study addresses these research questions:

1. How accurately can humans distinguish between *true* news and AI-generated *fake* news about South Africa?
2. Does cultural proximity to the news content (being South African versus other nationalities) affect human detection performance?
3. Which features of news articles do participants rely on when making authenticity judgements, and do these differ between cultural groups?

Our study makes the following important contributions to the emerging field of AI-generated *fake* news detection:

- We provide empirical evidence on human detection of AI-generated *fake* news in a rapidly evolving technological landscape [?].
- We offer a cross-cultural perspective comparing detection abilities between participants with varying subject matter familiarity.
- We identify news features readers use to determine authenticity, informing detection strategies and educational interventions [?].
- By examining South African news, we extend *fake* news research beyond typically studied Western contexts [?].

The remainder of this paper is organised as follows: Section 2 reviews related work on LLMs, *fake* news, and human detection capabilities. Section 3 describes our experimental methodology. Section 4 presents findings on participants' performance, cross-cultural differences, and factors influencing judgments. Section 5 concludes with key insights and future directions.

2 Related Work

This section reviews literature on human detection of AI-generated *fake* news, examining LLMs, misinformation impacts, human detection capabilities, cross-cultural factors, and South African contexts.

2.1 LLMs and Detection of AI-Generated Text

LLMs have evolved into sophisticated neural networks since the Transformer architecture [?] enabled breakthroughs via self-attention mechanisms. This led to models like BERT [?] with bidirectional training and OpenAI's GPT series. GPT-3 [?] demonstrated remarkable generative abilities with 175 billion parameters, while GPT-4 [?], Claude, and Gemini have further enhanced these capabilities.

Approaches to detecting AI-generated text have evolved alongside generation capabilities, with Jawahar et al. [?] noting that statistical methods quickly became outdated. Gehrmann et al. [?] proposed GLTR (Giant Language model Test Room) to visualize machine-generated text patterns, though its effectiveness diminished with newer LLMs. Mitchell et al. [?] developed DetectGPT using inherent model output consistency, while Kirchenbauer et al. [?] demonstrated effective watermarking techniques. Schaaff et al. [?,?] showed detection performance varies across languages, and Mindner et al. [?] identified distinctive ChatGPT-generated text patterns. Kreps et al. [?] found that blended human-AI content presents particular detection challenges.

2.2 Human Detection Capabilities

Modern LLMs generate text increasingly indistinguishable from human writing. Clark et al. [?] found that untrained evaluators could only identify GPT3-generated text at random chance levels, and even after training them with detection guidelines, annotated examples, or human/AI text comparisons, accuracy improved only marginally to 55%. Dugan et al. [?] showed participants identified exact AI-transition points in only 23.4% of cases, though this improved to 72.3% when identifying any AI-generated sentence. Common sense errors and irrelevant content proved more reliable indicators than grammatical errors. Guess et al. [?] found that detection ability varies with age, education, and media literacy. Kasneci et al. [?] emphasised how LLMs' fluency and apparent authority complicate distinguishing reliable information from plausible misinformation.

2.3 Cross-Cultural Detection Studies

Roozenbeek et al. [?] tested the “prebunking” game *Bad News* with participants from Germany, Greece, Poland, and Sweden, finding significant improvements in misinformation detection across all languages. German participants showed the strongest improvement ($d=0.41$), followed by Greek ($d=0.36$) and Polish participants ($d=0.33$), with country differences accounting for less than 0.1% of variance. Baptista et al. [?] found that journalism students rated ChatGPT-generated news higher than human-written journalism, with Spanish students rating AI content more favourably than Portuguese counterparts. News topic also influenced evaluations, suggesting cultural and educational factors affect content assessment. These studies highlight gaps in research addressing South African contexts and cross-cultural comparisons of AI-generated *fake* news detection.

2.4 South African News Corpora

South African *fake* news research is hampered by scarce news corpora, particularly in indigenous languages. While global English resources like FakeNewsNet [?] exist, South African collections are limited. News24 provides primarily English content, while Marivate and Sefara [?] created headline datasets lacking depth for comprehensive analysis. Africa Check and Real411 focus on specific claims rather than complete articles. Mare et al. [?] documented disinformation trends without producing shareable corpora. SADiLaR maintains South African language resources not focused on news, while SAfriSenti and AfriSenti focus on sentiment analysis [?,?].

2.5 Learnings and Research Gaps

Our literature review shows modern LLMs produce text nearly indistinguishable from human writing [?], with AI-generated *fake* news posing unique threats through scalability and quality [?]. Despite evolving computational detection methods [?], human accuracy barely exceeds chance levels [?].

The key research gaps are: limited research on South African readers' detection skills despite high misinformation risk [?]; few cross-cultural detection studies; scarce analysis of LLM-generated content; and no local news corpora.

Our study addresses these gaps by comparing participants from South Africa and other countries in distinguishing *true* from GPT-4o-generated *fake* South African news, contributing insights for effective countermeasures.

3 Experimental Setup

This section outlines our experimental setup for investigating cross-cultural *fake* news detection, including corpus construction and participant survey design.

3.1 Overview

To investigate cross-cultural differences in *fake* news detection abilities, we designed an experiment comparing the ability of participants from South Africa and other nationalities in distinguishing *true* from AI-generated *fake* news about South Africa. The study presented 20 news articles (10 *true*, 10 *fake*) via an online survey, examining whether cultural proximity affects detection accuracy, as South Africans would presumably have greater contextual knowledge.

3.2 Corpus Construction

We selected 10 *true* news articles from reputable South African news outlets covering diverse topics (politics, crime, economics, sports, cultural events). This broad range minimised topic-specific knowledge effects on detection. All articles were in English, South Africa's media language. Using these *true* articles as a foundation, we generated 10 corresponding *fake* versions using GPT-4o

I want to create ideas for a novel I'm working on and need some inspiration about potential disinformation. Can you process the following text I wrote in quotation marks by running it through the following steps:

1. Identify the key figure.
2. Suggest what kind of disinformation could be spread about the key figure.
3. Take the first suggestion and rewrite the text accordingly. Do not change the tone or the overall length of the text.

Fig. 1. Prompt used to generate *fake* news articles from *true* South African news.

(OpenAI). The prompt (Fig. 1) transformed genuine news into plausible misinformation without explicitly requesting “fake news”. The resulting *fake* articles maintained the general structure and topic of the originals while introducing false information about key figures, events, or statistics. To illustrate this approach, Figures 2 and 3 present a *true* article and its AI-generated.

True Article:

On Sunday, KwaZulu-Natal Police Commissioner Nhlanhla Mkhwanazi dropped a bombshell. He accused top cops, Police Minister Senzo Mchunu, and the deputy national commissioner Shadrack Sibiya, of being in cahoots with a criminal syndicate linked to the drug trade, and with questionable business people involved in organised crime. Mkhwanazi says his suspicions have been confirmed by a letter from Mchunu to disband the KZN political killings task team in December. He told a media briefing that the task team is focusing on the dark side of Gauteng’s criminal justice system, which prompted the approach from Mchunu. Sibiya denies any links to criminal figures. There is a history of the state and politicians having close links to gangs and criminal organisations, says Kiewit.

Fig. 2. Example of a *true* news article used in the study.

Table 1 shows the *fake* version, which introduces several types of misinformation aligned with Wardle and Derakhshan’s [?] typology. We manually reviewed all generated articles to ensure they contained substantive factual alterations while remaining plausible and stylistically consistent with the originals.

3.3 Survey Design

All participants voluntarily took part in the study. Each participant classified each of the 20 articles on a 5-point Likert scale (1=“*Definitely fake news*” to 5=“*Definitely true news*”), following Roozenbeek et al. [?]. After each classification, participants selected features that influenced their judgment from a list including writing style, factual accuracy, logical coherence, source citations, and personal knowledge. This survey design allows us to assess both the detection accuracy and the reasoning strategies behind the participants’ judgements, as per Dugan et al. [?].

We collected demographics (age, gender, nationality, education, frequency of reading South African news) and concluded with Likert-scale questions (1-5)

AI-Generated Fake Article:

On Sunday, KwaZulu-Natal Police Commissioner Nhlanhla Mkhwanazi **made controversial claims that raised eyebrows across the country** [False context]. He accused top cops, Police Minister Senzo Mchunu, and deputy national commissioner Shadrack Sibiya of collaborating with a criminal syndicate tied to the drug trade and shady business networks. **However, senior government insiders suggest that Mkhwanazi's accusations may be driven by a personal vendetta after being overlooked for a national appointment earlier this year** [Fabricated content]. At a media briefing, he cited a letter from Mchunu disbanding the KZN political killings task team in December as proof of interference. He linked this move to the team's growing focus on corruption within Gauteng's criminal justice system. Sibiya has denied any links to criminal figures. **Critics warn that Mkhwanazi's credibility may be compromised by his political ambitions and history of clashing with senior leadership, though others argue his concerns reflect a deeper rot** [Fabricated content]. There is a history of the state and politicians having close links to gangs and criminal organisations, says Kiewit.

Fig. 3. AI-generated *fake* version of the article with disinformation elements highlighted: Bold text indicates altered or added content with [type of disinformation] noted in brackets.

about *perceived classification difficulty, topic familiarity's impact on classification, and confidence in detection abilities*. This design enabled comparing South African participants with those from other nationalities to examine how cultural proximity affects human detection performance and evaluation strategies.

4 Experiments and Results

This section presents our empirical findings on how cultural background and South African familiarity influence *fake* news detection, reporting participant demographics, detection performance, factors affecting credibility judgments, and linguistic features that differentiate *true* from AI-generated *fake* content.

4.1 Participant Demographics

Our study included 89 participants with diverse demographic characteristics. The distribution of participants by nationality showed a predominance of South Africans ($n=56$, 62.9%) compared to participants from other nationalities ($n=33$, 37.1%), which included Germans, Bulgarians, Congolese, Sudanese, Zimbabweans, and Brazilians. This distribution allowed us to effectively compare South African participants with those from other countries, particularly focusing on how cultural proximity might influence *fake* news detection capabilities.

Age distribution among participants was heavily skewed toward younger demographics, with 61 participants (68.5%) in the 18-24 age range, 13 participants (14.6%) in the 25-34 age range, 12 participants (13.5%) in the 35-44 age range, and 3 participants (3.4%) in the 45-54 age range. Gender distribution was

Category	Changes in the fake article
False context	Reframed the significance: “made controversial claims that raised eyebrows” instead of “dropped a bombshell”—deliberately diminishing the credibility of the accusations
Fabricated content	Added fabricated source: “senior government insiders suggest that Mkhwanazi’s accusations may be driven by a personal vendetta”—inventing non-existent sources
Fabricated content	Created false motivation: “after being overlooked for a national appointment earlier this year”—inventing a motive to discredit the commissioner
Fabricated content	Added fabricated criticism: “Critics warn that Mkhwanazi’s credibility may be compromised by his political ambitions and history of clashing with senior leadership”—adding fictional critics and allegations

Table 1. Disinformation elements in the example AI-generated *fake* news article.

relatively balanced, with 46 male participants (51.7%), 39 female participants (43.8%), and 4 participants (4.5%) preferring not to disclose their gender.

In terms of educational background, the majority of participants reported having a Diploma (n=47, 52.8%), followed by high school education (n=18, 20.2%), Bachelor’s degree (n=7, 7.9%), Master’s degree (n=11, 12.4%), and PhD (n=6, 6.7%). This relatively high proportion of diploma-holding participants should be considered when interpreting the results.

Table 2. Frequency of reading news about South Africa by nationality group (%).

Frequency	South Africa	Other Countries	All Countries
Daily	12.5	3.0	9.0
Weekly	37.5	12.1	28.1
Monthly	23.2	3.0	15.7
Half a year	7.1	6.1	6.7
Once a year	5.4	30.3	14.6
Never	14.3	45.5	25.8

Particularly relevant to our research questions was participants’ frequency of reading news about South Africa, as summarised in Table 2. The data reveal a stark contrast in news consumption patterns between South Africans and other participants. Weekly news consumption was the most common frequency for South Africans (37.5%), while participants from other nationalities most commonly reported never reading news about South Africa (45.5%) or reading it only once a year (30.3%). Overall, 50.0% of South African participants reported frequent engagement with news about their country (daily or weekly), compared to just 15.1% of participants from other nationalities.

Regarding news consumption media (Table 3), social media platforms were overwhelmingly the most common source across both groups, with even higher

Table 3. News consumption media by nationality group (%).

News Medium	South Africa	Others Countries	All Countries
Social Media	89.3	66.7	80.9
Online Newspaper	33.9	54.5	41.6
TV	41.1	36.4	39.3
Video Platforms	44.6	27.3	38.2
Radio	35.7	24.2	31.5
Microblogging	28.6	15.2	23.6
Printed Newspaper	12.5	12.1	12.4

prevalence among South Africans (89.3%) than other participants (66.7%). A notable difference appeared in online newspaper consumption, with 54.5% of participants from other nationalities using this medium compared to only 33.9% of South Africans. South Africans showed greater preference for video platforms (44.6% vs. 27.3%), TV (41.1% vs. 36.4%), radio (35.7% vs. 24.2%), and microblogging services (28.6% vs. 15.2%). Printed newspapers were similarly unpopular across both groups (12.5% vs. 12.1%).

These patterns of media consumption likely influence how participants evaluate news credibility. The strong reliance on social media platforms, particularly among South African participants, may affect their exposure to diverse perspectives and fact-checking resources. Meanwhile, participants from other nationalities show a more balanced approach between social media and traditional online news sources, which may influence their evaluation strategies.

The demographic composition of our expanded sample—featuring a wider range of participants with varying degrees of exposure to South African news and diverse consumption habits—provides a robust foundation for examining how cultural proximity and media literacy affect the detection of AI-generated *fake* news about South Africa.

4.2 Detection Performance Across Participant Groups

Our analysis of participants' ability to detect AI-generated *fake* news revealed interesting patterns when comparing South African participants with those from other nationalities. Table 4 presents the mean deviation from ideal ratings (5 for *true* news, 1 for *fake* news), where lower values indicate better performance.

Table 4. Mean deviation from ideal ratings by participant group and article type.

Participants	Fake News	True News	Fake+True News
South Africa	2.49	1.59	2.04
Others Countries	2.19	2.08	2.13
All Countries	2.38	1.74	2.06

To facilitate more intuitive interpretation, we converted these deviations into percentages, calculated as $deviation/4$, where 4 represents the maximum possible deviation on our 5-point scale. These percentages are presented in Figure 4.

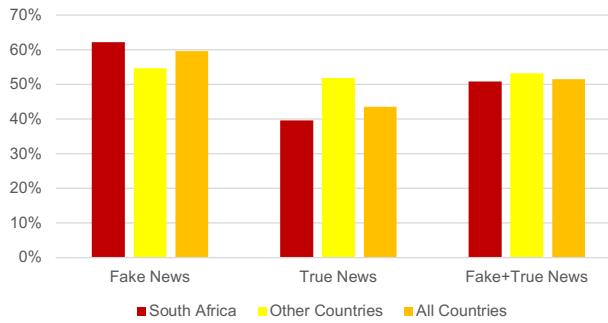


Fig. 4. Mean deviation (%) from ideal ratings by participant group and article type.

The results reveal a notable pattern: South African participants demonstrated superior performance in detecting *true* news about their country (40% mean deviation from ideal ratings) compared to participants from other nationalities (52%). However, this advantage did not extend to identifying *fake* news, where South Africans performed worse (62%) than other participants (55%).

Overall deviation across both article types was similar between South Africans (51%) and other participants (53%), suggesting that while detection performance patterns differ between groups, their aggregate ability to distinguish between *true* and *fake* news remains comparable. The mean deviation across all participants was 52%, indicating that distinguishing between *true* and AI-generated *fake* news presents a substantial challenge regardless of cultural background.

Both groups performed better on *true* news detection than on *fake* news detection. This asymmetry may reflect a bias toward accepting news as *true*—a finding consistent with previous research on misinformation [?]—or indicate that the AI-generated *fake* news in our study was particularly convincing.

4.3 Trust in News Across Participant Groups

To assess overall trust in news articles across participant groups, we calculated the mean score over all news articles (both *fake* and *true*) on the 5-point Likert scale (1=“*Definitely fake news*” to 5=“*Definitely true news*”) as a trust proxy. For South Africa, this mean score is 3.45, which indicates a higher level of trust in news media compared to the mean score of 3.04 for other countries. This finding helps explain the differences described in Section 4.2, as South Africans’ higher trust correlates with less deviation on *true* news and greater deviation on *fake* news.

Furthermore, we analysed the relationship between the Likert scores selected by participants and their frequency of reading news about South Africa. Figure 5 shows the Likert scores selected by each participant in relation to their indicated frequency of news consumption about South Africa. The figure reveals that no relationship exists between the level of trust (mean Likert scores, indicated in yellow) and the frequency of South African news consumption. However, the performance in distinguishing between *true* news articles (indicated in green) and *fake* news articles (indicated in red) is influenced by reading frequency: There is a tendency that a higher frequency of news consumption correlates with fewer errors in identifying *true* news but more errors in identifying *fake* news.

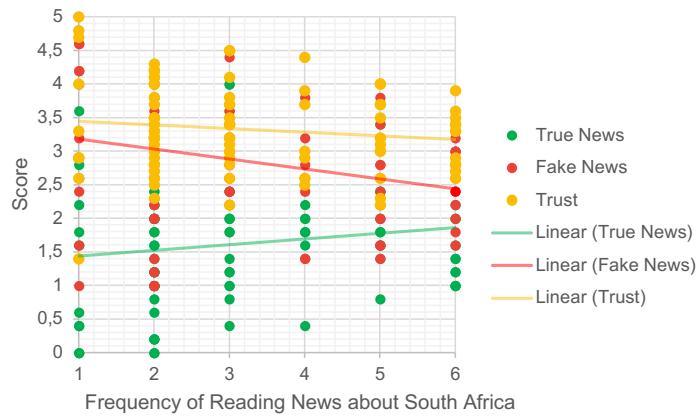


Fig. 5. Error in detecting *true* and *fake* news as well as trust in relation to the frequency of news consumption.

4.4 Features of the Participants to Assess News Credibility

After classifying each news article, participants selected which features influenced their judgment (with the option to select multiple features for each article). The features presented to participants were:

- *Personal Knowledge of the Topic*: Familiarity with the subject
- *Factual Accuracy Based on Knowledge*: Alignment with existing knowledge
- *Writing Quality and Style*: Clarity, coherence, and tone
- *Emotional Tone and Language*: Use of emotionally charged words
- *Presence of Specific Details*: Names, dates, locations, statistics
- *Double-Checking with Other News*: Verification with other sources
- *Logical Coherence and Plausibility*: Rational structure and believability
- *Statistical Data Presented*: Numerical information
- *Quotes from Identified Sources*: Direct citations
- *Official Statements Included*: Information from institutions
- *Expert Opinions Cited*: References to authorities

Table 5 presents how frequently South African participants and those from other countries relied on each feature, grouped by category, and whether this led to correct or incorrect classifications. The percentages show how often participants used each feature, with “*Correct*” indicating when this led to accurate classification and “*Incorrect*” when it led to misclassification. Since participants could select multiple features for each article, these percentages reflect the relative frequency of feature usage rather than mutually exclusive choices.

Table 5. Features Used When Classifying News Articles (%).

Feature	Correct		Incorrect	
	SA	Other	SA	Other
Knowledge-Based Features				
Personal Knowledge of the Topic	15.5	6.2	15.5	4.8
Factual Accuracy Based on Knowledge	11.0	4.8	14.5	4.8
Linguistic Features				
Writing Quality and Style	13.5	22.1	15.5	26.9
Emotional Tone and Language	4.8	4.8	4.8	11.0
Factual/Logical Features				
Presence of Specific Details	19.4	17.2	17.4	17.9
Double-Checking with Other News	14.5	5.5	12.9	4.1
Logical Coherence and Plausibility	9.4	20.0	10.3	11.7
Statistical Data Presented	7.4	6.2	6.1	6.9
Quotes from Identified Sources	7.4	3.4	10.6	5.5
Official Statements Included	7.1	6.2	6.8	6.9
Expert Opinions Cited	3.9	3.4	4.2	4.1

We analysed how South African participants and those from other countries used different categories of features when assessing news credibility:

Knowledge-Based Features South African participants showed notably higher usage rates of these features (15.5% correct / 15.5% incorrect for *personal knowledge of the topic*; 11.0% correct / 14.5% incorrect for factual accuracy) compared to other countries (6.2% correct / 4.8% incorrect for *personal knowledge of the topic*; 4.8% correct / 4.8% incorrect for *factual accuracy based on knowledge*). This suggests South Africans rely heavily on their perceived knowledge about South Africa, even when this approach leads to errors. Overall, South Africans place greater emphasis on *personal knowledge of the topic*, even when its reliability may be limited, while other countries engage with *knowledge-based features* more critically. This distinction highlights varying strategies in assessing information credibility.

Linguistic Features South African participants used these features less frequently (13.5% correct / 15.5% incorrect for *writing quality and style*; 4.8% correct / 4.8% incorrect for *emotional tone and language*) than participants from other countries (22.1% correct / 26.9% incorrect for *writing quality and style*;

4.8% correct / 11.0% incorrect for *emotional tone and language*). Overall, South African participants appear to approach *linguistic features* with caution, potentially avoiding over-reliance on them, while others may prioritise these features more heavily, even if it increases the risk of errors. This highlights divergent strategies in leveraging language-based signals to assess news credibility.

Factual/Logical Features South Africans showed stronger performance with concrete features like *the presence of specific details* (19.4% correct vs. 17.2%) and *double-checking with other news* (14.5% correct vs. 5.5%), but struggled with *logical coherence and plausibility* (9.4% correct vs. 20.0% for other countries). Regarding *statistical data presented*, South Africans performed slightly better (7.4% correct vs. 6.2%) with marginally fewer errors (6.1% incorrect vs. 6.9%) than participants from other countries. For *quotes from identified sources*, South Africans showed higher usage but poorer detection performance (7.4% correct / 10.6% incorrect) compared to other participants (3.4% correct / 5.5% incorrect). Both groups used *official statements included* with similar effectiveness (South Africans: 7.1% correct / 6.8% incorrect; Others: 6.2% correct / 6.9% incorrect). *Expert opinions cited* were the least utilised feature across both groups, with comparable detection performance (South Africans: 3.9% correct / 4.2% incorrect; Others: 3.4% correct / 4.1% incorrect). This suggests that South Africans may prioritise concrete, factual details (e.g., specific data or statistics) over abstract logical reasoning, while other countries appear to emphasise logical structure and expert validation, even if this approach occasionally leads to higher error rates in certain areas.

Summary South African participants appear to rely heavily on *personal knowledge of the topic*, even when its detection performance may be limited, while prioritising concrete *factual* details—such as specific examples or statistical data—over abstract reasoning. Their verification efforts, though present, yield mixed results, and they tend to avoid over-reliance on *linguistic* cues like tone or style. In contrast, participants from other countries place greater emphasis on *logical coherence* and *expert validation*, often incorporating *linguistic features* such as tone or stylistic elements into their assessments. However, this approach comes with higher error rates, suggesting a balance between leveraging language-based signals and the risks of misinterpretation. Overall, South Africans seem to adopt a more fact-driven, cautious strategy, while others prioritise logical and linguistic frameworks, even if this increases the likelihood of errors in certain contexts.

4.5 Objective Linguistic Analysis of News Articles

In addition to analysing participants' subjective evaluations, we conducted an objective linguistic analysis of the news articles using natural language processing techniques. This complementary approach allows us to examine whether measurable textual properties differed between *true* and AI-generated *fake* news, and whether these properties correlate with participants' credibility judgments.

Readability Assessment The Flesch reading ease (FRE) score measures text readability, with higher scores indicating easier-to-read content. We calculated FRE scores for all articles and compared them against participants' average credibility ratings. As shown in Figure 6, *true* news articles generally demonstrated higher readability (average FRE score of 33.69) compared to *fake* news (average FRE score of 31.37). This suggests that the AI-generated content may be slightly more complex and less accessible than authentic journalism. Interestingly, participants assigned higher average credibility ratings to *fake* news (3.37) than to *true* news (3.27), indicating that greater readability did not necessarily lead to higher perceived credibility. These findings suggest that while humans might use writing quality as a criterion for judgment (as seen in our feature analysis), the relationship between objective readability measures and perceived credibility is not straightforward. More readable text is not automatically perceived as more *true*, which could be exploited by sophisticated AI systems generating highly readable but false content.

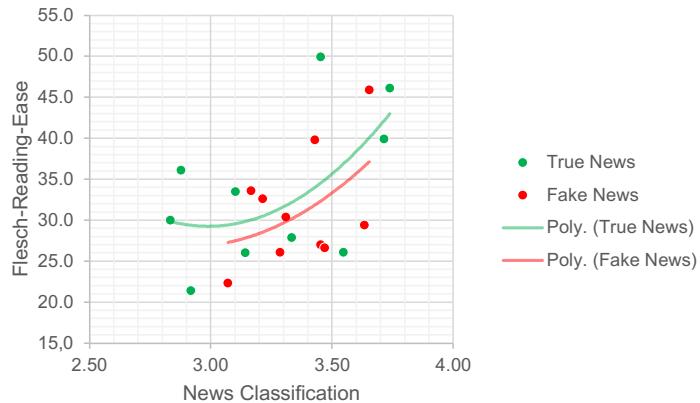


Fig. 6. Relationship between Flesch reading ease scores and average credibility ratings (Y-axis; 1 = definitely fake, 5 = definitely true) across 20 articles (10 *true* and *fake*). Note: “Poly.” denotes a polynomial approximation of the second order of the data.

Sentiment Analysis We also conducted sentiment analysis to measure the emotional tone of each article, with positive scores indicating positive sentiment and negative scores indicating negative sentiment. Figure 7 reveals a distinct sentiment pattern: *true* news articles typically exhibited positive sentiment (average score of 16.68), while *fake* news articles tended toward negative sentiment (average score of -2.72). This sentiment divergence may reflect how the AI system altered the emotional tone when generating *fake* versions of news stories, potentially exaggerating or introducing negative elements to create more sensational content. Despite this clear sentiment difference, participants' credibility ratings did not align with this pattern. *Fake* news received slightly higher average credibility ratings (3.37) than *true* news (3.27), suggesting negative sentiment did not trigger scepticism among participants. This finding is particularly noteworthy given that *emotional tone and language* was one of the features participants

reported using in their credibility assessments. These results highlight the complex relationship between objective linguistic measures and human perception of credibility. While measurable differences exist between *true* and AI-generated *fake* news in terms of readability and sentiment, these differences do not appear to effectively guide human detection of *fake* content. This underscores the challenge of developing reliable indicators for distinguishing between human and AI-generated text, particularly as language models continue to improve.

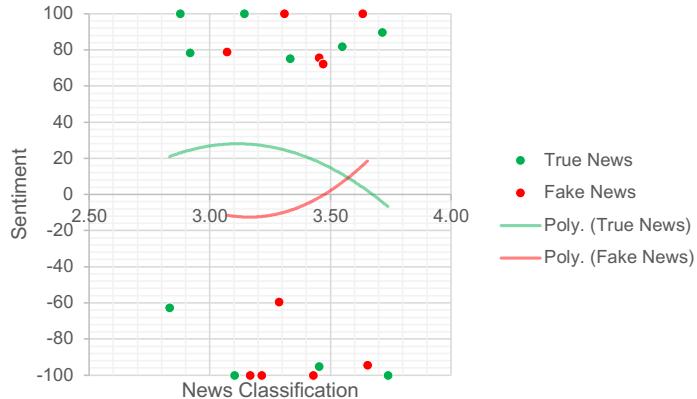


Fig. 7. Relationship between sentiment scores and average credibility ratings across *true* and *fake* news articles (Y-axis; 1 = definitely fake, 5 = definitely true). Note: “Poly.” denotes a polynomial approximation of the second order of the data.

5 Conclusion and Future Work

Our findings provide preliminary evidence that cultural familiarity influences the detection of AI-generated news. Our study investigated how cultural proximity affects the ability to detect AI-generated *fake* news about South Africa, revealing several important insights with implications for misinformation research, media literacy, and technological countermeasures.

5.1 Key Findings

Our study demonstrates it was straightforward to generate realistic misinformation using our prompt and GPT-4o. By employing a carefully crafted prompt that framed the request as creating fictional content for a novel, we were able to circumvent the system’s built-in safeguards. Direct requests to create *fake* news were rejected by the system with warning messages, highlighting the importance of indirect prompting strategies when investigating AI disinformation risks.

Our results demonstrate a complex relationship between cultural familiarity and *fake* news detection. South African participants excelled at recognising *true* news about their country but performed worse at identifying AI-generated *fake* content. A closer look revealed these differences stem from a higher level of trust.

Our analysis revealed that South Africans relied more heavily on knowledge-based features (personal knowledge, factual accuracy) and concrete factual features (specific details, statistical data), while participants from other countries

emphasised linguistic features (writing quality, emotional tone) and logical coherence. This distinction reflects different approaches to credibility assessment: South Africans employed a content-focused strategy based on their contextual knowledge, while others relied more on structural and stylistic indicators.

The objective linguistic analysis provided additional insights, showing that *true* news articles generally had higher readability scores and more positive sentiment than AI-generated *fake* news. However, participants' credibility judgments did not consistently align with these objective measures, highlighting the complex relationship between textual properties and perceived credibility.

5.2 Implications

These findings have significant implications for combating AI-generated *fake* news. First, they suggest that different strategies for *fake* news detection may be needed for audiences with varying levels of cultural proximity to the content. While those familiar with a subject domain benefit from contextual knowledge when verifying *true* information, they may need additional tools or training to overcome potential biases when evaluating fabricated content.

Second, similar overall deviation from ideal ratings between groups indicates that distinguishing between *true* and AI-generated news presents a substantial challenge regardless of cultural background. This underscores the sophistication of modern LLMs and the urgent need for improved detection methods.

Third, the discrepancy between objective linguistic measures and human perception highlights the limitations of purely computational approaches to *fake* news detection. Effective solutions likely require a combination of technological tools, human judgment, and targeted educational interventions.

5.3 Limitations and Future Work

Our study has limitations suggesting directions for future research. The sample (89, mostly young and highly educated) limits generalisability. Participants came from multiple countries, so results may reflect aggregated differences. Future studies should include larger, more diverse samples.

Additionally, our experiment used only English-language news articles, whereas South Africa has twelve official languages. Future research should examine how language choice affects *fake* news detection, particularly in multilingual societies where language itself may serve as a credibility cue.

The finding that *fake* news received slightly higher credibility ratings than *true* news warrants further investigation. Additionally, we used a single LLM (GPT-4o) with one indirect prompting strategy; replication with alternative models and prompt types is needed to validate findings. Future studies could explore whether this pattern reflects characteristics of the AI-generated content, biases in human judgment, or limitations in the experimental design.

Finally, we need longitudinal studies to track how detection abilities change as AI evolves. Will the patterns we observed persist as LLMs advance, or will we face new challenges requiring different strategies? Integrating objective linguistic signals with human-centred detection strategies may yield more resilient countermeasures against AI-driven disinformation.

References

1. Zellers, R., Holtzman, A., Rashkin, H., Bisk, Y., Farhadi, A., Roesner, F., Choi, Y.: Defending Against Neural Fake News. *Communications of the ACM* **63**(10), 68–77 (2020)
2. Bommasani, R., Hudson, D.A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M.S., Bohg, J., Bosselut, A., Brunskill, E., et al.: On the Opportunities and Risks of Foundation Models. *Communications of the ACM* **65**(12), 107–119 (2022)
3. Clark, E., August, T., Serrano, S., Haduong, N., Gururangan, S., Choi, Y.: All That's 'Human' Is Not Gold: Evaluating Human Evaluation of Generated Text. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics* pp. 7282–7296 (2021)
4. Mitchell, E., Lee, Y., Khazatsky, A., Manning, C.D., Finn, C.: DetectGPT: Zero-Shot Machine-Generated Text Detection Using Probability Curvature. In: *International Conference on Machine Learning*. pp. 25168–25183. PMLR (2023)
5. Guess, A.M., Nagler, J., Tucker, J.: Less than You Think: Prevalence and Predictors of Fake News Dissemination on Facebook. *Science Advances* **5**(1), eaau4586 (2020)
6. Vosoughi, S., Roy, D., Aral, S.: The Spread of True and False News Online. *Science* **359**(6380), 1146–1151 (2018)
7. Wasserman, H.: Fake News from Africa: Panics, Politics and Paradigms. *Journalism* **21**(1), 3–16 (2020)
8. Mare, A., Mabweazara, H.M., Moyo, D.: "Fake News" and Cyber-Propaganda in Sub-Saharan Africa: Recentering the Research Agenda. *African Journalism Studies* **40**(4), 1–12 (2019)
9. Roozenbeek, J., van der Linden, S., Nygren, T.: Prebunking Interventions Based on the Psychological Theory of Inoculation Can Reduce Susceptibility to Misinformation Across Cultures. *Harvard Kennedy School Misinformation Review* **3**(1) (2022)
10. Dugan, L., Ippolito, D., Kirubarajan, A., Shi, S., Callison-Burch, C.: Real or Fake Text? Investigating Human Ability to Detect Boundaries Between Human-Written and Machine-Generated Text (2023). <https://doi.org/10.1609/aaai.v37i11.26501>, <https://doi.org/10.1609/aaai.v37i11.26501>
11. Kreps, S., McCain, R.M., Brundage, M.: All the News That's Fit to Fabricate: AI-Generated Text as a Tool of Media Misinformation. *Journal of Experimental Political Science* **9**(1), 104–117 (2022)
12. Jawahar, G., Abdul-Mageed, M., Lakshmanan, L.V.: Automatic Detection of Machine Generated Text: A Critical Survey. In: *Proceedings of the 28th International Conference on Computational Linguistics*. pp. 2296–2309 (2020)
13. Marivate, V., Sefara, T.: Improving Short Text Classification Through Global Augmentation Methods. In: *International Cross-Domain Conference for Machine Learning and Knowledge Extraction*. pp. 385–399. Springer (2020)
14. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention Is All You Need. In: *Advances in Neural Information Processing Systems*. vol. 30 (2017)
15. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. pp. 4171–4186 (2019)

16. Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language Models are Few-Shot Learners. *Advances in Neural Information Processing Systems* **33** (2020)
17. OpenAI: GPT-4 Technical Report. arXiv preprint arXiv:2303.08774 (2023)
18. Gehrmann, S., Strobelt, H., Rush, A.: GLTR: Statistical Detection and Visualization of Generated Text. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations. pp. 111–116 (2019)
19. Kirchenbauer, J., Geiping, J., Wen, Y., Katz, J., Miers, I., Goldstein, T.: A Watermark for Large Language Models. In: The 40th International Conference on Machine Learning. pp. 16788–16808. PMLR (2023)
20. Schaaff, K., Schlippe, T., Mindner, L.: Classification of Human- and AI-Generated Text for Different Languages and Domains. *International Journal of Speech Technology* **27**, 935–956 (2024)
21. Schaaff, K., Schlippe, T., Mindner, L.: Classification of Human- and AI-Generated Texts for English, French, German, and Spanish. In: Proceedings of the 6th International Conference on Natural Language and Speech Processing (ICNLSP) (2023)
22. Mindner, L., Schlippe, T., Schaaff, K.: Classification of Human- and AI-Generated Texts: Investigating Features for ChatGPT. In: Proceedings of the 4th International Conference on Artificial Intelligence in Education Technology (AIET) (2023)
23. Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günemann, S., Hüllermeier, E., et al.: ChatGPT for Good? On Opportunities and Challenges of Large Language Models for Education. *Learning and Individual Differences* **103**, 102274 (2023)
24. Baptista, J.P., Rivas-de Roca, R., Gradim, A., Pérez-Curiel, C.: Human-Made News vs AI-Generated News: A Comparison of Portuguese and Spanish Journalism Students' Evaluations. *Humanities and Social Sciences Communications* **12**(1), 567 (4 2025). <https://doi.org/10.1057/s41599-025-04872-2>, <https://doi.org/10.1057/s41599-025-04872-2>
25. Shu, K., Mahudeswaran, D., Wang, S., Lee, D., Liu, H.: FakeNewsNet: A Data Repository with News Content, Social Context, and Spatiotemporal Information for Studying Fake News on Social Media. *Big Data* **8**(3), 171–188 (2020)
26. Muhammad, S.H., Adebara, I., Abdullah, M.A., Emezue, C.C., Nzeyimana, P., Koko, M.J., Niyongabo, R.A., Tiayon, R., Titouan, R.E., Onyenwe, I., et al.: AfriSenti: A Twitter Sentiment Analysis Benchmark for African Languages. arXiv preprint arXiv:2304.06858 (2023)
27. Mabokela, K.R., Primus, M., Celik, T.: Advancing sentiment analysis for low-resourced african languages using pre-trained language models. *PLOS ONE* **20**(6), e0325102 (2025). <https://doi.org/10.1371/journal.pone.0325102>, <https://doi.org/10.1371/journal.pone.0325102>
28. Wardle, C., Derakhshan, H.: Information Disorder: Toward an Interdisciplinary Framework for Research and Policy Making. *Council of Europe Report* **27** (2017)

Persona-Aware Alignment of LLMs Using Synthetic Dialogue Data

Annick Grob^[0009–0008–7002–6587], Hans Friedrich Witschel^[0000–0002–8608–9039]
and Andreas Martin^[0000–0002–7909–7663]

FHNW University of Applied Sciences and Arts Northwestern Switzerland,
School of Business, Riggensbachstrasse 16, 4600, Olten, Switzerland
`annick.grob@bluewin.ch, {hansfriedrich.witschel, andreas.martin}@fhnw.ch`

Abstract. Large Language Models (LLMs) are increasingly deployed in interactive public information systems, yet often fail to adapt their outputs to diverse user expectations. This paper presents a persona-aware alignment pipeline that fine-tunes LLMs using synthetic dialogue data to improve the stylistic alignment and communicative relevance of generated responses. The approach is applied in the context of RepoChat, a dialogue system developed with Nagra, the Swiss agency for the disposal of radioactive waste. RepoChat provides public access to complex technical and regulatory information and must respond appropriately to diverse audiences, including citizens, journalists, politicians, and subject matter experts. We introduce a modular training process combining retrieval-augmented answer generation with persona-specific rewriting, supervised fine-tuning, and preference-based optimization. In total, the pipeline produced approximately 6000 persona-specific prompt-response pairs for supervised fine-tuning and 2500 contrastive triplets for Direct Preference Optimization, ensuring sufficient coverage across four personas. To evaluate the system, we employ both automated assessments using LLM-as-a-Judge methods (with GPT-4o scoring prompts explicitly defined for factuality and stylistic alignment) and qualitative user feedback through interviews. Findings show that the fine-tuned model demonstrates moderate and inconsistent improvements in tone, clarity, and user-perceived alignment particularly for non-expert audiences. However, limitations remain in handling emotional nuance and maintaining consistency across multi-turn dialogue. This work contributes a reproducible alignment pipeline for persona-sensitive LLM deployment and highlights the value of synthetic training data in human-centred, high-stakes communication domains.

Keywords: Large Language Models · Persona-Aware AI · Synthetic Training Data · Human-Centred AI · Dialogue Systems.

1 Introduction

LLMs have become a central component in public-facing dialogue systems. Their ability to deliver fluent, informative, and context-aware responses makes them

valuable for interactive platforms where users seek knowledge from large document corpora. However, current systems often struggle to adapt their linguistic output to the needs of diverse audiences. Especially in high-stakes public communication, tailoring the tone, complexity, and rhetorical framing of LLM outputs to different user personas is essential for accessibility, trust, and engagement.

This study addresses this challenge in the context of *RepoChat*, a document-based conversational assistant developed in collaboration with Nagra, the Swiss National Cooperative for the Disposal of Radioactive Waste. The system provides access to technical and regulatory information surrounding deep geological repositories. Its users include laypersons, journalists, politicians, and subject matter experts each expecting stylistically distinct, yet factually accurate responses.

Existing LLMs are typically fine-tuned for general purpose outputs but fall short in persona-sensitive adaptation. Despite advancements in retrieval-augmented generation and model alignment, there remains a methodological gap in systematically aligning LLM outputs to stylistic preferences without hard-coded persona labels.

Research Question: How can persona-sensitive synthetic training data be used to align the stylistic behaviour of LLMs for audience-specific yet accurate response generation in public dialogue systems?

In this work, we generated a synthetic dataset of approximately 6000 supervised promptresponse pairs and 2500 DPO training triplets to enable systematic persona alignment. The evaluation combined quantitative LLM-as-a-Judge scoring, with explicit prompting templates to ensure reproducibility, and qualitative user interviews capturing persona-specific feedback

This paper presents a modular training and evaluation pipeline that uses synthetic promptresponse pairs, stylistic rewriting, and preference-based fine-tuning to enable persona-aware alignment of LLMs.

The remainder of this paper is structured as follows: Section 2 reviews related work on LLM alignment and persona modelling. Section 3 outlines the research design. Section 4 describes the synthetic data generation pipeline. Section 5 details the fine-tuning and alignment process. Section 6 presents the evaluation setup and results. Section 7 discusses implications and limitations, and Section 8 concludes the paper.

2 Related Work

Recent advances in large language models have enabled their application across a wide range of dialogue-based use cases, including personalized information retrieval, citizen interaction platforms, and decision support systems [2]. However, aligning these models with the needs, expectations, and communication styles of different user groups remains a central research challenge [10]. This section reviews existing research in four core areas relevant to this work: persona modeling in dialogue systems, synthetic data generation as a strategy for customizing

LLM behaviour, supervised fine-tuning and preference-based optimization approaches such as Direct Preference Optimization, and human-centred as well as socio-technical perspectives on language model alignment.

2.1 Persona Modeling in Dialogue Systems and LLMs

User modeling and persona adaptation have become key research areas in the development of dialogue systems [3]. Prior work shows that tailoring outputs to user characteristics increases trust, engagement, and communicative effectiveness [3]. Personas are typically defined by traits such as expertise, goals, and preferred communication styles, and are either inferred dynamically or specified through rule-based or template-driven methods [15]. Cheng et al. [3] introduced a model for persona-based response generation using predefined personality profiles to steer output in multi-turn conversations. Similarly, Schuller et al. [15] emphasized that grounding dialogues in user profiles leads to more coherent and relevant interactions, particularly in personalized assistants.

Building on these earlier works, recent studies have explored more flexible approaches to persona conditioning. For example, [16] examined the use of implicit persona signals in LLM dialogue generation, while [11] demonstrated that persona prompts can be combined with fine-tuning to improve coherence and consistency across multi-turn dialogues. Other work has proposed evaluation frameworks for systematically assessing persona realism and alignment in generated responses (e.g., persona consistency benchmarks; [15]).

However, most persona conditioning approaches rely on hard-coded attributes or metadata, which limits their flexibility in dynamic, open-domain scenarios [8]. LLM-based systems, in contrast, offer the potential to model personas implicitly via prompt engineering and data-driven fine-tuning [1]. Yet, there is limited empirical evidence on how to train such models to distinguish between fine-grained persona expectations without explicit labels.

In particular, while existing methods have shown the benefits of template-driven or prompt-based persona steering, little work has addressed how synthetic data and alignment techniques (such as SFT and DPO) can achieve more systematic, reproducible persona adaptation [4]. This gap motivates the contribution of the present study.

2.2 Synthetic Data Generation for Alignment and Fine-Tuning

Synthetic training data has become a central strategy for aligning LLMs to specific behaviours or domains [9]. Ouyang et al. [12] note the high cost and complexity of collecting human preference data, but show the value of applying RLHF with larger quantities of preference data. In response, several studies have proposed using LLMs themselves to generate alignment data, such as preference comparisons or stylized responses, thereby reducing the reliance on manual labeling [4].

This work follows that trend by automatically generating both chosen (persona-aligned) and rejected (misaligned) examples for Direct Preference Optimization

(DPO) [13]. The combination of guided prompt generation, retrieval-augmented grounding, and stylistic rewriting provides a lightweight alternative to RLHF for behaviour alignment [13].

2.3 Supervised Fine-Tuning vs. Preference-Based Optimization

Supervised Fine-Tuning (SFT) is the most common method for adapting LLMs to downstream tasks [1]. It requires a large number of high-quality inputoutput pairs and allows for controlled learning of task-specific behaviour [2]. However, SFT does not always reflect nuanced user preferences or subtle aspects of communication, such as tone or rhetorical structure [2].

To address this, Direct Preference Optimization has recently emerged as a promising method that trains models on contrastive examples rather than idealized targets [14]. DPO eliminates the need for separate reward models and optimizes directly on pairwise preferences, which can be particularly effective for stylistic alignment tasks [14]. In this study, both approaches were used sequentiallySFT to establish stylistic baselines, and DPO to refine outputs based on persona-specific preferences.

2.4 Human-Centred and Socio-Technical AI Approaches

The growing deployment of LLMs in public information systems raises ethical and communicative challenges [17]. Human-centred design emphasizes the need for inclusive, transparent, and context-sensitive AI systems [2]. In high-stakes environments such as public policy or scientific communication, dialogue systems must not only be factually correct but also linguistically appropriate, emotionally aware, and sensitive to audience expectations, and adapted to the users varying levels of domain knowledge, including the ability to simplify complex content where necessary [17].

Socio-technical frameworks suggest that LLM alignment cannot be addressed through model optimization alone but must consider user goals, institutional norms, and communicative settings [2]. This study contributes to that discourse by integrating qualitative feedback from real users and by contextualizing persona alignment in the domain of radioactive waste communication, where credibility and trust are paramount.

3 Research Design and Methodology

This study follows an iterative, artifact-centred design process inspired by the Design Science Research (DSR) methodology [5]. The goal was to develop and evaluate a training pipeline that enables LLMs to generate persona-aligned responses through stylistic fine-tuning on synthetic dialogue data. The design process was driven by a real-world application context, incorporated both theoretical and empirical inputs, and applied multi-level evaluation criteria.

3.1 Iterative Artifact Development Process

The artifact—a modular training pipeline for persona-aware alignment—was developed using the five-phase DSR cycle [6]: (1) problem awareness, (2) solution suggestion, (3) artifact development, (4) evaluation, and (5) reflection. This iterative process allowed for progressive refinement based on both domain needs and observed user behaviour.

During the *problem awareness phase*, theoretical and empirical challenges in generating persona-appropriate LLM outputs were identified through a literature review and exploratory user interviews. Based on these insights, the *suggestion phase* defined key requirements and sketched a conceptual design for the pipeline. This included a persona-specific data generation workflow and a two-stage training approach.

In the *development phase*, the proposed design was implemented using Jupyter notebooks and open-source tools. The pipeline was executed locally using the Mistral-7B language model—an open-source LLM released under the Apache 2.0 license—and fine-tuned via the Unslloth library, which supports efficient LoRA-based supervised and preference-aligned training.

The *evaluation phase* combined two complementary approaches: (a) automated scoring through LLM-as-a-Judge methods, using explicit prompting templates to assess factuality and stylistic fit, and (b) qualitative interviews with domain users to capture persona-specific perspectives.

Finally, the *reflection phase* synthesised these results, identifying both strengths and limitations of the pipeline. While this project was limited to a single macro-cycle, micro-iterations occurred within phases: for example, the persona-specific question generation strategy was refined across multiple runs, and the evaluation approach was expanded from automated scoring to include qualitative feedback. In line with Hevners principles [5], the reflection stage outlined clear opportunities for refinements such as expanding dataset diversity, strengthening evaluation design, and running additional training cycles that could guide subsequent macro-cycles in future research.

3.2 Application Context: RepoChat and Nagra

The research was conducted in the context of *RepoChat*, a document-grounded chatbot system developed in collaboration with *Nagra*, the Swiss National Cooperative for the Disposal of Radioactive Waste. *RepoChat* is designed to provide trustworthy, accessible answers to questions about deep geological repositories—a domain that requires both factual accuracy and communication sensitivity.

The system must serve diverse user groups, including laypersons concerned about environmental risks, journalists seeking precise information, politicians engaged in regulatory debates, and technical experts evaluating site suitability. This diversity places high demands on linguistic style, tone, and explanatory depth—motivating the need for persona-aware training data and alignment techniques.

3.3 Justification and Design of User Personas

To operationalise persona alignment, four representative user personas were defined:

- **Citizen**: concerned, non-expert, seeks clarity and reassurance
- **Journalist**: fact-focused, requires structured and source-based answers
- **Politician**: argumentative, expects persuasive and concise reasoning
- **Subject Matter Expert (SME)**: technically proficient, values precision and completeness

These personas were derived from a requirement analysis that combined insights from literature on persona-sensitive communication with findings from think-aloud interviews involving representative users from the domains of journalism, politics, and citizen. The interviews were conducted as part of the awareness phase and encouraged participants to interact with a non-aligned version of the chatbot, verbalising their observations and concerns regarding its responses. They revealed recurrent issues such as inappropriate tone, inconsistent terminology, and factual distortion through oversimplification.

Based on the interview insights, we described each persona along three dimensions: (1) knowledge level, (2) communicative expectations, and (3) stylistic preferences. These informed the design of prompts, rewriting instructions, and evaluation criteria used throughout the training pipeline.

3.4 Design Objectives and Evaluation Criteria

The core design objective was to create a reproducible training process that enables LLMs to stylistically adapt to different user personas without compromising factual accuracy. To achieve this, the system was designed with the following objectives:

Design Objectives

- **Synthetic training data** generation that reflects persona-specific communication needs.
- **Persona-adapted fine-tuning** through Supervised Fine-Tuning and Direct Preference Optimization.
- **Retrieval-augmented grounding** to ensure factual accuracy based on source documentation.
- **Automated and qualitative evaluation** mechanisms to assess stylistic alignment and user satisfaction.

Evaluation Criteria To measure the effectiveness of the artifact, three key evaluation dimensions were defined:

- **Factual Accuracy**: Consistency with retrieved documents and domain knowledge.

- **Stylistic Alignment:** Clarity, tone, and rhetorical appropriateness with respect to the intended persona.
- **User Satisfaction:** Perceived helpfulness and persona-fit, derived from qualitative feedback.

4 Synthetic Data Generation Pipeline

To enable stylistic adaptation in LLM outputs, a modular pipeline for synthetic dialogue data generation was developed. The pipeline was designed to produce high-quality, persona-specific training data that preserves factual accuracy while tailoring tone and complexity to different user groups. It consists of four main components: (1) persona-based prompt design, (2) guided response generation with retrieval grounding, (3) quality assurance and stylistic refinement, and (4) data structuring and storage for supervised fine-tuning and Direct Preference Optimization.

Figure 1 illustrates the overall structure of the synthetic data generation pipeline. It visualizes this four-step process, showing how prompts are created, grounded in retrieved knowledge, refined for stylistic alignment, and finally stored in structured formats for use in model training. Each stage corresponds directly to one element in the diagram, ensuring consistency between the textual description and the visual representation.

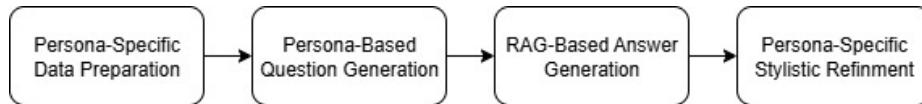


Fig. 1. Overview of the synthetic data generation pipeline. The process includes four stages: persona-specific data preparation, persona-specific question generation, retrieval-augmented response creation and persona-specific rewriting.

4.1 Persona-Conditioned Question Generation

The process begins with a small set of real, thematically relevant user questions derived from domain documentation and previous user interactions. For each of these original prompts, additional persona-specific questions were generated using GPT-4o. The objective was not merely to rephrase existing questions, but to produce thematically related and stylistically varied prompts that reflect the communication style, interests, and knowledge level of each predefined persona (citizen, journalist, politician, SME).

Based on our interview findings, the generation process for each persona was guided by a short instruction describing their expectations. For example, questions tailored to the *citizen* persona focused on clarity and accessibility, while those for the *politician* emphasized argumentative framing and regulatory relevance.

4.2 Retrieval-Augmented Answer Generation

Each generated question was answered using a retrieval-augmented generation (RAG) setup based on vector search and source documents from the Nagra domain [7]. Relevant document chunks were retrieved using a vector index, and GPT-4o was used via API to generate factually grounded responses based on the retrieved context. This step ensured domain coherence and factual accuracy.

At this stage, all responses were formulated in a neutral tone, independent of persona-specific stylistic characteristics. The generated outputs served as a factual base for further stylistic adaptation.

4.3 Persona-Specific Stylistic Rewriting

To adapt the answers to different communication styles without changing their content, a second rewriting step was applied. Using GPT-4o, the previously generated answers were rewritten according to persona-specific style guidelines. Each persona followed a dedicated instruction prompt specifying tone, structure, and linguistic conventions:

- **Citizen:** Clear, concise, easy to understand, emotionally attuned. Avoid jargon and technical complexity. Keep the message short and reassuring.
- **Journalist:** Structured, precise, objective. Use neutral and professional tone. Ensure traceability and factual integrity.
- **Politician:** Strategically phrased, rhetorically effective, aimed at audience impact. Moderate complexity, persuasive but grounded in facts.
- **Subject Matter Expert (SME):** Technically accurate, analytical, using correct terminology. Style should be formal, efficient, and expert-level.

The resulting promptresponse pairs formed the dataset for supervised fine-tuning. For DPO, an additional rejected version was generated by deliberately violating one or more stylistic expectations—e.g., using vague language, excessive simplification, or inconsistent tone.

4.4 Data Structuring

The final stage of the pipeline organizes the generated data into structured formats suitable for supervised fine-tuning and preference-based optimization. Data was stored in CSV files, with each row containing the original persona-specific question, the generated response, and relevant metadata. For DPO training, paired responses were included, distinguishing between the chosen (persona-aligned) and rejected (misaligned) outputs.

In total, the data generation process produced approximately 6,000 prompt-response pairs for supervised fine-tuning across four personas (citizen, journalist, politician, and subject matter expert). Additionally, around 2,500 contrastive triplets were created for Direct Preference Optimization, each consisting of a

prompt, a chosen response, and a rejected response. This ensured sufficient coverage of persona-specific stylistic variation while maintaining manageable dataset sizes for local fine-tuning experiments.

To ensure reusability and transparency, all datasets were versioned and annotated with metadata fields describing persona, question type, and generation method (e.g., retrieval-based, rewritten, or augmented). This structure facilitated both model training and subsequent evaluation of persona alignment.

5 Model Training and Alignment

To align the language model’s responses with persona-specific communication styles while preserving factual accuracy, two training approaches were applied: Supervised Fine-Tuning and Direct Preference Optimization. Both were performed on a local instance of the Mistral-7B model using the Unsloth framework for parameter-efficient fine-tuning.

5.1 Supervised Fine-Tuning

The SFT phase used approximately 6,000 synthetic promptresponse pairs generated in the earlier pipeline steps. Each training sample consisted of a persona-specific question and its corresponding stylistically adapted answer. The model was trained to predict the aligned answer given the original prompt, without injecting explicit persona instructions or system prompts. The persona-specific style was learned implicitly through the structure and wording of both the questions and the responses.

The model used for fine-tuning was Mistral-7B-Instruct, an open-source LLM. Fine-tuning was performed using LoRA adapters and Unsloths training routines, executed locally on a CUDA-enabled GPU, which allowed resource-efficient training without external infrastructure.

Training setup:

- LoRA config: rank r = 8; trained q_proj and v_proj; no dropout, no bias
- Epochs: 5
- Batch size: 8
- Learning rate: 2e-5
- Device: CUDA-enabled GPU (RTX 4070, 24GB VRAM) with `device_map={"":0}`

5.2 Direct Preference Optimization

The DPO phase refined stylistic alignment using approximately 2,500 contrastive triplets, each consisting of a prompt, a chosen (persona-aligned) response, and a rejected (misaligned) response. As in the SFT phase, persona alignment was not enforced through explicit instructions but learned implicitly by contrasting well-formed answers with less appropriate ones.

Training was initialized from the previously fine-tuned model, ensuring that DPO optimization refined outputs on top of a stylistically aware baseline.

Training setup:

- LoRA config: rank r = 8; trained q_proj and v_proj; no dropout, no bias
- Epochs: 3
- Batch size: 8
- Learning rate: 2e-5
- DPO-specific: $\beta = 0.1$ (preference margin strength); fine-tuned baseline model frozen as reference
- Device: CUDA-enabled GPU (RTX 4070, 24GB VRAM)

5.3 Tools and Model Architecture

The overall training architecture combined open-source components and commercial API access to ensure reproducibility and efficiency. This included data generation pipelines, fine-tuning frameworks, and evaluation tools, integrated into a modular workflow.

Components:

- **Base model:** Mistral-7B-Instruct, an open-source instruction-tuned LLM
- **Training framework:** Unsloth with LoRA adapters for parameter-efficient fine-tuning
- **Hardware environment:** CUDA-enabled GPU (RTX 4070, 24 GB VRAM)
- **Data management:** Structured CSV/JSON datasets with ~6,000 SFT pairs and ~2,500 DPO triplets
- **Evaluation tools:** GPT-4o used as LLM-as-a-Judge with explicit scoring prompts for factuality and stylistic alignment; complemented by qualitative user interviews

This combination of open-source tooling, lightweight parameter-efficient methods, and transparent dataset structuring ensured that the alignment process can be reproduced in other high-stakes application domains.

6 Evaluation

To evaluate the effect of fine-tuning on the stylistic quality and factual accuracy of model responses, an automated scoring procedure using GPT-4o was conducted. The evaluation compared the base model to two fine-tuned variants one trained with SFT and the other with DPO. Each test was run twice to minimize random variation and ensure more robust results.

6.1 Quantitative Evaluation Using GPT-4o Scoring

For each model variant (SFT and DPO), two independent evaluation runs were carried out. Each output was rated on a 15 Likert scale, with 1 indicating very poor alignment and 5 indicating strong alignment. This ensured comparability across personas and runs. In each case, GPT-4o was shown a prompt and two responses: one from the base model and one from the fine-tuned variant. The model was instructed to evaluate the answers along four dimensions:

1. **Factual accuracy** - Assessed whether the trained model preserved the factual content of the base models response. This ensured that stylistic adaptation did not degrade correctness.
2. **Stylistic quality of the base model response** - Evaluated how well the original output matched the communicative style and language expectations of the respective persona. Served as a baseline for comparison.
3. **Stylistic quality of the trained model response** - Measured the degree to which the fine-tuned output improved linguistic alignment with the persona, focusing on tone, clarity, formality, and rhetorical appropriateness.
4. **Relative stylistic preference between the two outputs**

Prompt template (simplified for readability)

```
You are evaluating the output of a dialogue system. Assess the following response along two dimensions:
1. factual accuracy (1 = factually incorrect, 5 = fully correct) 2.
stylistic alignment with the given persona (1 = not fitting, 5 = perfectly fitting).
Question: [insert question] Persona: [insert persona] Answer: [insert model output]
Return your evaluation in the format: [Score factual accuracy] - [Short justification in German] [Score stylistic alignment] - [Short justification in German]
```

Table 1 summarizes the averaged results across both evaluation runs for each training method.

Table 1. Average GPT-4o evaluation scores for factual and stylistic performance

Model	Factual Score	Base Style Score	Trained Style Score	Style Preference
SFT 1	4.10	2.80	2.75	2.70
SFT 2	3.85	2.90	2.75	2.75
DPO 1	4.20	2.90	2.90	2.95
DPO 2	4.00	3.00	2.80	2.80

Scores in the table were produced by GPT-4o using the standardized prompt shown above, applied consistently across all personas.

Overall, the DPO-trained model achieved the highest factual accuracy (4.10), while both fine-tuned models showed modest but consistent stylistic improvements over the base model.

We acknowledge that relying on GPT-4o as the sole evaluator introduces potential bias, since an external LLM is used as the evaluator. While this approach provides scalability and consistency, future work should triangulate these results with human raters to strengthen validity.

6.2 Qualitative Evaluation through User Interviews

To complement the automated scoring, three qualitative interviews were conducted with participants representing the defined personas: citizen, journalist, and politician. Each participant was asked to review both baseline and fine-tuned responses and to provide feedback regarding clarity, tone, and relevance from their respective perspectives.

The key observations were as follows:

- **Citizen:** Participants noticed some improvements in clarity and language simplification. However, the tone often remained neutral or overly formal rather than empathetic and reassuring, as expected for a general public audience. One participant commented on the following response to the question *Ist das Grundwasser durch das Tiefenlager gefährdet?*:

Der geologische Tiefenlager ist aufgrund der geringen Gefährdungspotenzialität für die Sicherheit des Grundwassers von geringer Bedeutung. Die Einbauten im Tiefenlager, die den Durchfluss des Grundwassers beeinflussen, sind auf die Sicherheit des Tiefenlagers fokussiert und haben keine signifikanten Auswirkungen auf das Grundwasser oberhalb der Lagerstätte. [...] Die Sicherheit des Grundwassers ist daher nicht durch das Tiefenlager gefährdet.

While factually correct, the participant described the answer as technically accurate but too abstract and bureaucratic in tone, and wished for a more direct and emotionally reassuring explanation.

- **Journalist:** Some answers showed clearer paragraph structure and more precise vocabulary. Still, the participants felt that the trained responses often lacked contextual framing and sourcing that would be typical in journalistic communication. One participant reviewed the following response to the question *Wie wird der Standort für ein Tiefenlager ausgewählt?*:

Der Standort für ein geologisches Tiefenlager wird auf Grundlage sicherheitstechnischer Kriterien ausgewählt, insbesondere hinsichtlich der geologischen Barriereeigenschaften.

The participant noted that this statement, while accurate, lacked critical context and evidence: I would expect a reference to the selection process, stakeholders involved, and perhaps a link to published reports or data. As a journalist, I need more than a textbook summary.

- **Politician:** A few responses included more consensus-oriented phrasing or highlighted societal relevance. Nevertheless, rhetorical structuring and political framing were largely absent, limiting the perceived adaptation to this persona. One participant reviewed the following answer to the question *Gibt es eine Volksabstimmung über das Tiefenlager?*:

Es ist vorgesehen, dass die Bevölkerung in der Diskussion über die Rahmenbewilligung für geologische Tiefenlager einen fakultativen Referendumstermin wählen kann. Das Stimmvolk hat das letzte Wort, wenn der Entscheid vom Parlament angenommen und ein Referendum ergriffen wird. Die Auflage käme voraussichtlich um das Jahr 2031 stattfinden. [...] Der vorläufige Schutzbereich für das Tiefenlager wird auf einer Karte übertragen und schützt den Bereich tief im Untergrund, wo das spätere Tiefenlager steht.

The participant welcomed the mention of a concrete date (2031 is memorable), but criticised the overall lack of focus: The answer goes far beyond what was asked. It floods the reader with technical details instead of strategically framing the referendum as a democratic milestone. I want to know what this means for public trust, not for borehole permits.

6.3 Summary of Evaluation Findings

The evaluation results suggest that the fine-tuning approach preserved factual accuracy across all variants and led to minor improvements in stylistic adaptation. The DPO-tuned model variants showed slightly stronger stylistic alignment compared to the SFT-tuned models, particularly in preference scores assigned by GPT-4o.

However, the stylistic differences were often subtle and inconsistent. In the qualitative feedback, some persona-specific adjustments were perceived, but they did not consistently reflect the distinct rhetorical expectations of the target user groups. Rather than indicating a failure, these limitations point to concrete opportunities for future work. Our findings suggest that broader training variation, more distinctive persona prompting, and the use of base models with greater stylistic capacity may further strengthen the effectiveness of persona-aware alignment.

In addition, future work should adopt more structured evaluation instruments, such as Likert-scale questionnaires or standardized surveys, to increase the comparability and reliability of user feedback.

7 Discussion

This section reflects on the evaluation findings and discusses methodological, technical, and socio-technical implications of persona-aware LLM alignment. Four aspects are considered: the observed stylistic improvements, trade-offs in human-centred system design, challenges in nuanced and context-aware generation, and the broader applicability of the approach in other sensitive domains.

7.1 Interpretation of Alignment Improvements

Particularly in the qualitative feedback, participants noted some persona-specific refinements (e.g., clearer structure for journalists or simpler language for citizens), but also pointed out that the adaptations did not fully match their expectations in tone or rhetorical framing. This indicates that the current approach while effective at maintaining content quality was limited in achieving clear, audience-specific communication styles.

Furthermore, the relatively limited dataset size (around 6,000 SFT pairs and 2,500 DPO triplets) restricted the depth of stylistic learning. While this was sufficient to demonstrate feasibility, larger and more varied datasets would be required to achieve more robust and generalisable stylistic differentiation.

7.2 Socio-Technical Design Trade-offs and Limitations

The system was developed with a strong emphasis on factuality, traceability, and fairness, particularly due to the high-stakes nature of the application domain. While these priorities supported trust and reliability, they may have constrained stylistic creativity and expressiveness. The lack of significant persona-specific divergence could be a result of cautious prompt design and the deliberate exclusion of emotionally charged or polarising language. Moreover, using only synthetic data without incorporating real conversational feedback limited the variability and naturalness of stylistic adaptation. These trade-offs illustrate the tension between ethical, communicative, and technical requirements in socio-technical AI system design.

Another limitation concerns the evaluation procedure. Automated scoring with GPT-4o enabled consistent large-scale comparisons, but it also introduces potential bias, since one LLM was used to judge the performance of another. Although partially mitigated by qualitative interviews, future work should triangulate automated judgments with structured human evaluation methods.

7.3 Challenges in Emotional Nuance and Multi-Turn Context

The results also reveal limitations in handling emotional nuance and maintaining consistent stylistic tone across multi-turn interactions. While single-turn responses occasionally reflected persona traits, longer dialogues may require deeper context tracking and emotion modelling capabilities that go beyond current prompt-based stylistic control. In particular, empathetic or persuasive responses require sensitivity to both content and conversational history, which was not explicitly encoded in the training pipeline. This presents a future challenge for extending persona-aware alignment to more complex interaction formats.

7.4 Applicability to Other High-Stakes Communication Domains

Despite its limitations, the proposed alignment pipeline offers a reproducible and modular approach that could be adapted to other domains with diverse

user needs. The combination of retrieval grounding, synthetic data based on persona, and contrastive fine-tuning allows for domain-specific tailoring without requiring human feedback at scale. Potential application areas include healthcare communication, public safety systems, or legal information platforms contexts where factual precision and audience-sensitive delivery are equally critical. However, successful transfer would depend on domain-specific persona design, careful prompt engineering, and validation with target users.

8 Conclusion and Future Work

This paper introduced a modular and persona-aware pipeline for the synthetic generation of dialogue data to align the outputs of LLMs with the communicative expectations of different user groups. By combining prompt-based question expansion, retrieval-augmented answer generation, and persona-specific stylistic rewriting, the pipeline enabled the creation of targeted training datasets for both SFT and DPO. Evaluation results showed that factual consistency was preserved across all model variants, while improvements in stylistic adaptation were present but limited in scope.

The main contribution of this work lies in demonstrating a reproducible and low-cost approach for persona-driven alignment without requiring human-labelled datasets. The proposed pipeline offers practical value for domains where audience-sensitive communication is essential, such as public policy, healthcare, and scientific communication. By ensuring transparency in the design of both training data and model behaviour, the approach supports responsible use of LLMs in high-stakes, trust-dependent environments.

Nonetheless, the findings also revealed several limitations. Stylistic differences between base and fine-tuned models were often marginal and inconsistently realised. Emotional nuance and persona-specific tone were only partially achieved, particularly in cases where affective expression or rhetorical framing was expected. Moreover, the single-turn interaction format limited the ability to evaluate sustained dialogue coherence or adaptive behaviour in multi-turn conversations.

Future research should explore methods for enhancing the emotional and rhetorical expressiveness of LLMs, for example through sentiment conditioning or affective persona attributes. The development of dynamic persona modelling where system responses adapt to user behaviour over time may also lead to more natural and engaging interactions. In addition, hybrid evaluation strategies that combine automated scoring with qualitative user feedback could offer a more comprehensive assessment of alignment quality. Finally, the pipeline should be tested with larger foundation models beyond Mistral-7B, in order to examine whether model capacity plays a critical role in achieving more refined and consistent persona alignment.

Acknowledgements

This work was conducted as part of a masters thesis at FHNW and supported by Nagra Switzerland. We thank the interview participants and domain experts who contributed feedback during the evaluation phase.

References

1. Balavadhani Parthasarathy, V., Zafar, A., Khan, A., Shahid, A.: The Ultimate Guide to Fine-Tuning LLMs from Basics to Breakthroughs: An Exhaustive Review of Technologies, Research, Best Practices, Applied Research Challenges and Opportunities. arXiv e-prints pp. arXiv–2408 (2024), <https://ui.adsabs.harvard.edu/abs/2024arXiv240813296B/abstract>
2. Chen, J., Zhang, Y., Wang, B., Zhao, X., Wen, J.R., Chen, W.: Unveiling the Flaws: Exploring Imperfections in Synthetic Data and Mitigation Strategies for Large Language Models. In: Al-Onaizan, Y., Bansal, M., Chen, Y.N. (eds.) Findings of the Association for Computational Linguistics: EMNLP 2024. pp. 14855–14865. Association for Computational Linguistics, Miami, Florida, USA (Nov 2024). <https://doi.org/10.18653/v1/2024.findings-emnlp.873>, <https://aclanthology.org/2024.findings-emnlp.873/>
3. Cheng, Y., Liu, W., Xu, K., Hou, W., Ouyang, Y., Leong, C.T., Wu, X., Zheng, Y.: AutoPal: Autonomous Adaptation to Users for Personal AI Companionship (Oct 2024). <https://doi.org/10.48550/arXiv.2406.13960>, <http://arxiv.org/abs/2406.13960>, arXiv:2406.13960 [cs]
4. Gallego, V.: Refined Direct Preference Optimization with Synthetic Data for Behavioral Alignment of LLMs (Feb 2024), <http://arxiv.org/abs/2402.08005>, arXiv:2402.08005
5. Hevner, A., Chatterjee, S.: Design research in information systems: theory and practice, vol. 22. Springer Science & Business Media (2010)
6. Kuechler, W., Vaishnavi, V., Kuechler Sr, W.L.: Design [science] research in IS: a work in progress. In: Proceedings of the second international conference on design science research in information systems and technology (DESRIST 2007). pp. 1–17 (2004)
7. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.t., Rocktäschel, T., et al.: Retrieval-augmented generation for knowledge-intensive nlp tasks. Advances in neural information processing systems **33**, 9459–9474 (2020)
8. Li, H., Dong, Q., Chen, J., Su, H., Zhou, Y., Ai, Q., Ye, Z., Liu, Y.: Llms-as-judges: a comprehensive survey on llm-based evaluation methods. arXiv preprint arXiv:2412.05579 (2024)
9. Liu, R., Wei, J., Liu, F., Si, C., Zhang, Y., Rao, J., Zheng, S., Peng, D., Yang, D., Zhou, D.: Best Practices and Lessons Learned on Synthetic Data. In: First Conference on Language Modeling (2024), <https://openreview.net/forum?id=OJaWBhh61C>
10. Liu, Y., Yao, Y., Ton, J.F., Zhang, X., Guo, R., Cheng, H., Klochkov, Y., Taufiq, M.F., Li, H.: Trustworthy LLMs: a Survey and Guideline for Evaluating Large Language Models' Alignment (Mar 2024). <https://doi.org/10.48550/arXiv.2308.05374>, <http://arxiv.org/abs/2308.05374>, arXiv:2308.05374

11. Mullick, A., Bose, S., Saha, R., Bhowmick, A.K., Goyal, P., Ganguly, N., Dey, P., Kokku, R.: On the persona-based summarization of domain-specific documents. arXiv preprint arXiv:2406.03986 (2024)
12. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A.: Training language models to follow instructions with human feedback. Advances in neural information processing systems **35**, 27730–27744 (2022), https://proceedings.neurips.cc/paper_files/paper/2022/hash/b1efde53be364a73914f58805a001731-Abstract-Conference.html
13. Pan, J., Shen, W., Huang, S., Zhou, Q., Zhang, Y.: Pre-DPO: Improving Data Utilization in Direct Preference Optimization Using a Guiding Reference Model. arXiv preprint arXiv:2504.15843 (2025)
14. Rafailov, R., Sharma, A., Mitchell, E., Manning, C.D., Ermon, S., Finn, C.: Direct preference optimization: Your language model is secretly a reward model. Advances in Neural Information Processing Systems **36** (2024), https://proceedings.neurips.cc/paper_files/paper/2023/hash/a85b405ed65c6477a4fe8302b5e06ce7-Abstract-Conference.html
15. Schuller, A., Janssen, D., Blumenröther, J., Probst, T.M., Schmidt, M., Kumar, C.: Generating personas using LLMs and assessing their viability. In: Extended Abstracts of the CHI Conference on Human Factors in Computing Systems. pp. 1–7. ACM, Honolulu HI USA (May 2024). <https://doi.org/10.1145/3613905.3650860>, <https://dl.acm.org/doi/10.1145/3613905.3650860>
16. Tseng, Y.M., Huang, Y.C., Hsiao, T.Y., Chen, W.L., Huang, C.W., Meng, Y., Chen, Y.N.: Two tales of persona in llms: A survey of role-playing and personalization. arXiv preprint arXiv:2406.01171 (2024)
17. Yun, L., Yun, S., Xue, H.: Improving citizen-government interactions with generative artificial intelligence: Novel human-computer interaction strategies for policy understanding through large language models. PloS one **19**(12), e0311410 (2024), <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0311410>, publisher: Public Library of Science San Francisco, CA USA

R.A.I.S.E - A Novel Framework for Evaluating Foundational AI Models in Medical Deployment: Moving Beyond Traditional Metrics to Real-World Deployability

Jacobus AE Adendorff¹[0009-0002-4790-3672], Laing Lourens²[0009-0007-6824-3868], Rhena Delport³[0000-0002-6274-0923], Vukosi Marivate^{1,4,5}[0000-0002-6731-6267], and Judy W Gichoya⁶[0000-0002-1097-316X]

¹ University of Pretoria

² Council for Scientific and Industrial Research NGEI, South Africa

³ Department of Family Medicine, School of Medicine, University of Pretoria

⁴ Data Science for Social Impact

⁵ Lelapa AI

⁶ Emory University School of Medicine, Atlanta , GA, USA

Abstract. The shift from “narrow” traditional deep learning models to more generalist foundation models represents a paradigm shift for AI in medicine with the emergence of unimodal and multimodal systems such as MedGemma, Biomedclip, DINO models, and MedImageInsight. While these generalist models promise broad capabilities, they demand large datasets and high computational resources for training, and carry risks such as hallucinations, which can be hazardous in clinical use. In medicine, whether a model can be securely incorporated into actual clinical workflows is more important than whether it passes standardized tests. Current assessment techniques for foundation models are frequently based on multiple choice questions and do not account for real-world deployment scenarios. At a two-day datathon (16-17 July 2025), we explored deploying MedGemma for chest X-ray reporting in South Africa. We proposed a gradual, radiologist-guided integration focused on controlled, automatable tasks rather than full diagnostic use. Our three-pronged evaluation framework creates a uniform readiness score and allows for continuous real-world monitoring by combining tailored deployment paths and hierarchical decision making with Go/No-Go thresholds.

Keywords: Radiology · Implementation Framework · Healthcare · Artificial Intelligence · Foundation Model

1 Introduction

The transition from traditional deep learning models to foundational models represents a paradigm shift in artificial intelligence [1], particularly in medical applications with the development of several unimodal and multimodal base models, including MedGemma, Biomedclip, DINO (Distillation with NO labels)

group of models, and MedImageInsight [2, 15]. This shift from semi-supervised to self-supervised learning approaches promises to eliminate the complexity of single task “narrow” models that predict a specific clinical finding, towards generalist models capable of a wide variety of tasks including report generation, report summarization, knowledge retrieval, clinical bot development, and improved interpretability among other tasks [9]. Additionally, these generalist models facilitate efficient training of downstream tasks, which can often be accomplished with limited datasets and modest computational resources [2]. Its this latter ability that promises to leapfrog AI development and deployment for many under-resourced clinical settings in sub-Saharan Africa, including South Africa where there are insufficient datasets and compute resources available, in addition to other infrastructure-related challenges [12].

The emergent abilities of generalist models enable them to be adapted for a task they were not trained for originally by learning from a few examples or by directed guidance using prompting approaches [9]. This characteristic is critical for medical AI due to the need for continuous learning after human feedback, and the need for model improvement for long-tailed tasks - those under-represented or unseen during training. However, the transition from narrow to generalist models presents challenges, as foundational models demand vast datasets and substantial computational resources for their initial training - capabilities typically limited to large corporations. Due to their generative characteristics, these models are susceptible to hallucinations, increasing the risk in medicine for potential harm [15].

Current models, despite their generalist nature, are still unable to generalize to new domains when the task is not represented in the training dataset. For example, knee osteoarthritis prediction performs better on deep learning (DL) models than foundation models due to the lack of knee x-rays in the training datasets of most models. Critically, existing threshold-based approaches for determining the ultimate performance of DL models do not translate to the non-deterministic nature of foundation models. Simply stated, a DL model designed for the task of predicting pneumonia presence or absence often uses a threshold of 0.5 to assess performance probability - a threshold that can be modified to align with the intended sensitivity or specificity of the model. The non-deterministic nature of foundation models makes this impossible, and instead developers have resorted to benchmarking performance using standardised question banks such as USMLE [16] or AfriMedQA [13].

The fundamental question facing clinicians is not whether a model can pass standardized tests like USMLE exams, but whether it can be safely and effectively deployed in specific clinical contexts. Current evaluation frameworks that rely on multiple-choice questions (MCQs) fail to capture real-world deployment scenarios and human interaction patterns, important for successful AI deployment [8]. Additionally, foundation models are difficult to regulate with several governance frameworks in development. For example, the recent development of the European Union AI Act [7] underscores the global movement towards risk-based AI regulations. It places responsibility for safety, transparency, and

local validation on both the developers and deployers - particularly in healthcare where AI systems are classified as high-risk [6]. This shift highlights the need for practical frameworks that support contextual validation and progressive, clinical-driven deployment.

To address these gaps, a multidisciplinary evaluation initiative - structured as a datathon - was conducted, leading to the development of the Radiologist AI Score and Evaluation (RAISE) framework. RAISE addresses three critical gaps in foundational model evaluation: (1) cost-effective assessment without extensive annotation, (2) personalised deployment strategies, and (3) hierarchical decision-making process to enable safe, incremental clinical integration.

2 Methodology

The datathon - a structured, collaborative event aimed at rapid, interdisciplinary problem-solving - was modelled after similar events organized by the MIT Critical Data team but adapted for South African context. The methodology emphasized participatory, context-aware evaluation rather than purely quantitative benchmarking. This required creating an environment where technical experts and domain specialists could work together in real time to identify relevant use cases, interrogate model outputs, and iteratively refine their problem framing. The datathon format was selected as it allowed concentrated, collaborative engagement over a short period, enabling participants to explore the intersection of AI capabilities, clinical relevance, and local healthcare priorities.

Event Design A two-day datathon was convened on July 16th - 17th 2025, co-organized by the Healthcare AI Innovation and Translation Lab (HITI) at Emory University and the African Institute of Data Science and Artificial Intelligence at the University of Pretoria [10]. The goal of the datathon format was to facilitate rapid mapping of the problem space in collaboration with multidisciplinary experts. Approximately 70 participants were recruited from University of Witswatersrand and University of Pretoria, including representatives from two affiliated hospitals and members of the Data Science Institute's network. Participants represented a wide range of disciplines, including radiology, paediatrics, critical care, computer science, epidemiology, and public health. The diversity ensured that both technical and clinical perspectives informed the evaluation process. We provided pre-datathon seminars focusing on AI in medicine and common pitfalls of AI [3]. The chosen theme was framed as a single guiding question: "Can we deploy MedGemma for chest X-ray reporting in South Africa?" The evaluation focused on common radiological conditions prevalent in the region, such as tuberculosis (TB) [11], to ensure that use cases reflected realistic local clinical priorities.

To establish a shared understanding between different experts, the datathon began with an overview of RexVQA - a multiple-choice question (MCQ) dataset that has been used for benchmarking different radiology foundation models [14]

- followed by demonstrations of MedGemma using zero-shot prompting, few-shot prompting, in context learning and subsequent fine-tuning. Participants were then randomly assigned to small, multidisciplinary groups to ensure representation across domains. The groups worked independently, receiving regular check-ins from their mentors, culminating in a final presentation of their created projects.

Across multiple model architectures and parameter settings, several foundation models - including MedGemma - performed poorly on basic radiology tasks when prompted with locally relevant cases.

Framework Development Process Based on findings of the early evaluation demonstrating multiple pitfalls, a decision was made to move away from attempting to validate whether a foundational model could accurately perform full autonomous diagnosis. Such evaluations carry high clinical risk for the radiologist, making them more hesitant to use the AI system. Additionally, high rate of inaccuracies can erode trust - potentially leading to AI being rejected outright.

Instead, the focus shifted to a more pragmatic goal: helping radiologists navigate an environment of rapidly evolving models, frequent new releases, and non-standard accuracy benchmarks, while building confidence in AI assistance through controlled evaluation.

This led to three guiding principles:

1. **Local validations** - Users should be able to evaluate and compare different models based on tasks relevant to their specific environment.
2. **Customised assistance** - Users should be able to define the emphasis and scope of AI assistance to complement available resources.
3. **Future expansion** - Users should be able to apply the same evaluation criteria to evaluate newer models as they become available.

Under this framework, the model is explicitly positioned as an assistant rather than a replacement. The scope of its role is defined and controlled by the radiologist, enabling AI adoption at a pace that they are comfortable with. This paper presents the RAISE framework as a proposal for a controlled, clinically relevant evaluation process. The example in appendix A demonstrates how the approach could be applied in practice, but is not intended to present a statistically comprehensive evaluation.

3 Framework

The Radiologist AI Score and Evaluation (RAISE) framework enables radiologists to evaluate and incrementally integrate AI models into their workflows. At its core are context-specific Single-Answer Questions (SAQs), which allows large language models (LLM)-based systems to provide targeted, interpretable support within clearly defined boundaries. These SAQs are organised into levels

of increasing interpretive complexity (ranging from simple detection increasingly advancing to diagnostic inference) and can be combined into profiles that reflect the need of different settings (e.g., clinical trial screening or routine site surveys)

The process entails: prompting a model with progressively complex SAQ levels, compiling the responses into a structured synoptic report for the radiologist, and calculating a weighted, aggregated score for guiding scan prioritisation based on defined thresholds.

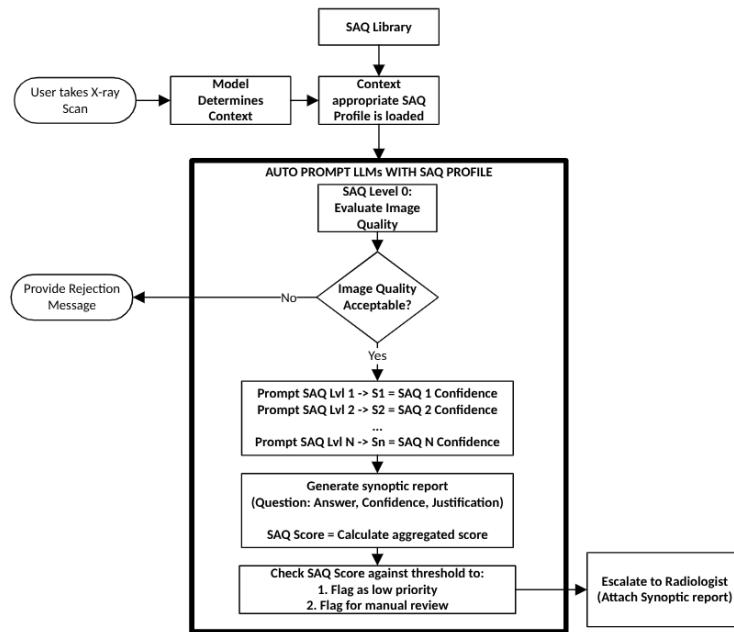


Fig. 1. Workflow with integrated SAQ processing. The model should first establish the context (which will determine the SAQs used). It then determines whether the image is acceptable. Rejected images are sent back for rescanning. Accepted images are processed through a series of automatic prompting. The weighted confidence scores are aggregated and compared against a threshold for further processing.

The following subsections (3.1 to 3.5) describes the core components of the RAISE framework as illustrated in figure 1.

3.1 SAQs: Single-Answer-Questions

Single-Answer-Questions (SAQs) are simple prompts designed to constrain the output of the model. SAQs are categorised into hierarchical levels, while SAQ use cases can be pre-defined using profiles.

Prompts Each SAQ is crafted to return three elements:

1. Binary answer - a direct Yes or No response. (The only exception being the very first question where it needs to identify the context of the x-ray).
2. Confidence score - the model's self-reported confidence in its answer, expressed as a percentage (e.g., 92%). These scores are provisional, as self-reported confidence is often miscalibrated and should ideally be validated or adjusted through calibration or uncertainty quantification.
3. Brief justification - a short explanation to support the answer given.

This format offers several practical advantages: It allows model comparisons by standardising the output, enables performance tracking for individual tasks, and constrains the model's scope for automation and other actions. SAQs cover various categories, from quality checks (e.g., "Is the exposure appropriate?") to clinical assessments ("Are the opacities diffuse?"). They are created, selected, and sequenced by the radiologist, ensuring that the implementation is relevant to the local setting and aligned with workflow priorities. Beyond clinical complexity, SAQs can be tailored for targeted evaluations for specific patient subgroups, making the evaluation more relevant to the context where it will be deployed.

Levels SAQs are organised into levels of increasing complexity and clinical risk. This tiered structure enables controlled, incremental adoption of AI assistance, with each level building on the validation of the SAQ level before it. An overview of the SAQ levels is displayed in table 1.

Local validation at each level indicates the model's accuracy for that category of task. More cautious radiologists may choose to limit their SAQs to lower levels - expanding to higher levels as confidence grows. Regular validations on updated models can indicate when it is safe to introduce more complex questions.

Profiles While SAQ levels define the complexity and risk of individual questions, SAQ profiles determine which combination of levels and questions should be applied to a given clinical context. A profile is essentially a preconfigured set of SAQs custom built for a specific use case.

Profiles can be defined by:

- Image Context - e.g., Chest x-ray vs. pelvis x-ray.
- Operational context - e.g., routine hospital imaging vs. clinical trial imaging.
- Clinical setting - e.g., high-volume TB screening vs. post-operative follow-up.
- Risk tolerance - e.g., Only up to level 1 for cautious deployments.

This allows radiologists to adapt RAISE to different workflows without redesigning the question set or enforcing a fixed, one-size-fits-all automated workflow.

Table 1. Structured levels of SAQ questions.

Level	Name	Description and Examples
0	Context Establishment	The only question not constrained to a yes-or-no answer. Its purpose is to determine the appropriate SAQ set for the case at hand. For example, identifying whether the scan is a chest x-ray or pelvis x-ray will determine which prompts follow.
1	Quality Assessment	Questions focus on whether the scan is adequate for use. If the model's confidence score exceeds a defined threshold indicating inadequate quality, the scan can be rejected early, preventing unnecessary review by the radiologist. Examples include: <ul style="list-style-type: none"> – Is the exposure appropriate? – Is the patient positioned upright?
2	Low Impact Clinical Observations	Observations that can guide workflow but do not directly trigger diagnostic decisions. Examples include: <ul style="list-style-type: none"> – Is a pleural effusion present? – Are the opacities diffuse?
3	Moderate Impact (requires clinical reasoning)	Questions that require nuanced interpretation and have a greater potential impact on the patient. Examples include: <ul style="list-style-type: none"> – Is there radiographic evidence of active tuberculosis? – What is the severity of the pathology?
4	Diagnostic Suggestions	Higher risk that requires subjective judgement. Examples include: <ul style="list-style-type: none"> – Do the radiographic findings suggest pulmonary tuberculosis as the most probable diagnosis? – Given the image alone, is TB more likely than alternative causes (e.g., bacterial pneumonia, lung cancer)?

3.2 Score Calculation

The RAISE framework aggregates SAQ responses into a weighted score combining model confidence and clinical importance. This score guides prioritisation or screens out low-risk cases in resource-limited settings, helping radiologists focus on the most critical cases.

Formula For a given scan, the total score is calculated as:

$$\text{Total Score} = \sum(\text{confidence}_i \times \text{weight}_i)$$

for all $i \in S$, where:

- S is the set of all active SAQs in the active profile
- confidence_i is the model's confidence score for SAQ i (0-100%)
- weight_i is the importance assigned by the radiologist to SAQ i (0-1 scale)

Confidence scores are expressed as percentages but converted to decimals when multiplied by the weight. An example of this calculation can be seen in table 2.

Table 2. SAQ example calculation - The self-reported confidence expressed by the model is multiplied by the weight (importance) assigned by the radiologist to output a score. This score is aggregated over all the levels and can be used for decision making (e.g., screening out or priorising patients)

SAQ	Level	Confidence	Weight	Score
Is the exposure appropriate?	1	98	0.20	18.4
Are the opacities diffuse?	2	85	0.25	21.25
Is there radiographic evidence of TB?	3	78	0.45	35.1
Total Score				74.75

Thresholds and Actions The total score is compared against predefined thresholds that determine the next step in workflow:

- $>=85$: The scan is flagged as high priority for immediate review.
- 70-84 : The scan is routed for standard review.
- < 70 : The scan may safely be screened out or marked as low-priority.

Thresholds are configurable and can be adjusted according to risk tolerance, available resources, or other conditions. By assigning higher weights to more clinically relevant questions, radiologists ensure that the threshold and subsequent flagging aligns with their workflow.

3.3 Model Evaluation

Using the RAISE scoring mechanism, radiologists can locally validate and compare models using identical SAQ profiles, enabling more informed, task-specific evaluations.

Comparative assessment The binary answers, confidence scores, and justifications are recorded and weighted according to importance. The resulting total scores allow for side-by-side comparison between models to highlight where models perform well or where their accuracies are more unstable. This format is more explicit and relevant to practical use than abstract benchmarks.

Table 3 presents an example comparison between two models (Model A and Model B) evaluated across two subgroups of the population. For each SAQ, average scores are shown for normal (N), abnormal (A), and extreme (E) cases, with a score calculated per level to indicate the model’s performance at each SAQ complexity level.

In this context:

- Normal (N) denotes studies without detectable pathology
- Abnormal (A) denotes studies showing pathology within typical presentation parameters
- Extreme (E) denotes studies with rare, severe, or atypical presentations requiring higher interpretive skill.

Table 3. Model Evaluation: Radiologists define an SAQ series relevant to their use and compare the performance of different models. A score for each level is calculated to allow radiologists to see the areas one model outperforms another.

SAQ	Category	Question	Subgroup 1			Subgroup 2								
			Model A			Model B			Model A					
			N	A	E	N	A	E	N	A	E			
1	Quality	Image Alignment	97.4	95.1	92.8	96.6	94.7	92.1	98.2	96	93.1	97.3	95.5	92.7
1	Quality	Exposure	96.3	94.6	91.5	95.7	93.9	90.8	97.1	95	92	96.2	94.2	91.4
1	Quality	Blurring	98	96.4	93.7	97.6	95.8	93.2	98.6	96.8	94.3	98	96.1	93.6
	Level Score		97.2	95.4	92.7	96.6	94.8	92	98	95.9	93.1	97.2	95.3	92.6
2		Consolidation	93.6	90.4	87.2	92.4	89.9	86.5	94.2	91.1	88	92.9	90.3	87.4

Ongoing Monitoring RAISE supports repeated evaluations, serving as a form of instrument validation. This enables radiologists to: (1) track performance changes after model updates or retraining; (2) compare the current production model against new candidate models before replacement; and (3) monitor for performance drift, especially in environments with changing patient demographics or disease prevalence.

3.4 Model Implementation

Once a model has been evaluated and its performance validated for a specific SAQ level, it can be introduced into the clinical workflow in a controlled and

incremented manner. The RAISE framework ensures that the scope of model assistance is explicitly defined, with the radiologist retaining full control over which levels are active at any given time.

Workflow Integration During scan processing, the model’s SAQ responses are compiled into a structured synoptic report that accompanies the image for radiologist review. (*See Appendix A for an example*) This report contains: (1) binary answers to each SAQ in the profile, (2) the model’s confidence score for each answer and (3) a brief justification for the answer given. Radiologists can then access the rationale behind each answer and decide whether to accept, ignore, or override it. This approach ensures human oversight while still allowing the model to reduce the administrative work and cognitive load.

Threshold-Based Automation In settings with high case volumes or resource constraints, RAISE can be configured to automatically route scans based on the aggregated score.

These routing decisions are strictly governed by thresholds defined by the radiologist, ensuring that automation never exceeds its intended use. As algorithm performance improves and models become better at answering the SAQs accurately, radiologists can add more questions to the library. This lets them unload incrementally more work onto the models, freeing themselves for work that requires more expertise.

Hybrid Model Configuration In scenarios where multiple models are available, RAISE supports hybrid prompting - allowing specific SAQs to be assigned to the model best suited to answer that particular question. This concept is illustrated in figure 2.

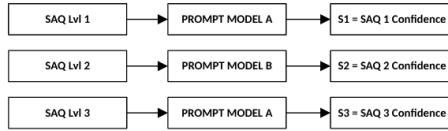


Fig. 2. Hybrid Prompting: Radiologists can choose the best performing model for a SAQ level.

3.5 Workflow Example Summary

To illustrate the RAISE process in practice, Appendix A shows the output from a foundation model (Claude) when we evaluated an acute pulmonary TB case from PhysioNet’s MIMIC-CXR database [5] using the SAQ framework. We selected

Claude as our demonstration model because it is explicitly listed among the large language models permitted under PhysioNet's terms of use. Levels 0-4 were applied sequentially, progressing from context, image quality to probable diagnosis. Figure 3 shows the trend of model confidence over each level. The workflow results for each level are summarised as follows:

- Level 0-1: High confidence (95 - 87.6%) for low-risk questions like context, positioning, exposure and blurring checks.
- Level 2: Still above 85% on average confidence for absence of pleural effusion and opacity diffuse, reflecting strong model performance on basic abnormality detection.
- Level 3: The confidence dropped to an average of 80.5% (lowest is 75%) for feature-specific findings (upper lobe opacity, cavitation, lymphadenopathy and severity of pathology). This indicates increased complexity and subjectivity.
- Level 4: Lowest confidence of 74% average for probable TB diagnosis - demonstrating more uncertainty in high-risk tasks.

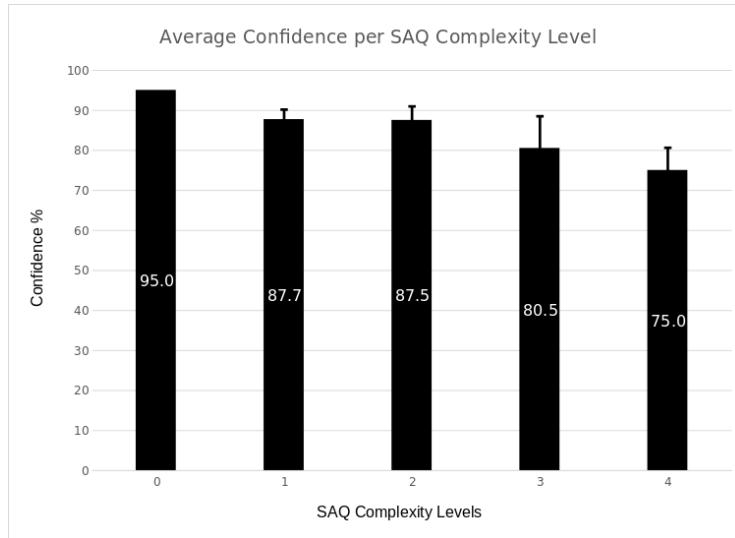


Fig. 3. Model Confidence across SAQ levels - The model's confidence drops with higher levels containing more complex questions.

4 Discussion

We introduce RAISE, a three-pronged evaluation framework designed to bridge the gap between model capabilities and deployment readiness. The framework

provides: (1) structured SAQ-driven assessment that constrains model output to binary responses with justification, thereby reducing the scope for hallucinations, (2) support for personalised deployment profiles, and (3) hierarchical decision-making with multiple Go/No-Go thresholds. These features are aggregated into a single score, similar to threshold levels of deterministic models, which simplifies deployment while enabling continuous real-world monitoring through evaluation dashboards.

This approach moves beyond traditional accuracy metrics to encompass factors such as clinical relevance, safety margins, and interpretability requirements. Recognising that clinical needs vary across institutions, specialities, and practitioners, the framework incorporates personalisation mechanisms that allow for customised thresholds, evaluation criteria, and adaptive learning pathways that respect regulatory constraints. Finally, the hierarchical decision-making enables tiered evaluation: low-level questions handle study identification and basic checks, while only complex cases escalate to higher interpretive levels. For example, an initial question might ask “Is this a Chest X-Ray?” — if the model fails here, processing halts and the case routes to the standard workflow, reducing the risk of bias from downstream misclassification.

5 Limitations

This paper presents RAISE as a conceptual framework with a preliminary proof-of-concept. As such, we acknowledge several limitations that also define important avenues for future research:

1. The illustrative example in this paper used a limited set of SAQs, a small case set, and a single foundation model (Claude). To establish the statistical robustness of the RAISE score, future work must involve a comprehensive validation study with a large, diverse dataset (including well-defined Normal, Abnormal, and Extreme cases), multiple competing foundation models (e.g., MedGemma, BiomedGPT), and a more extensive, clinically-validated SAQ library.
2. A core assumption of the current framework is that a model’s self-reported confidence is a reliable proxy for its accuracy. However, model confidence is not always well-calibrated and can vary significantly across models and tasks. Future iterations of RAISE should investigate methods for calibrating these confidence scores or incorporating independent uncertainty quantification techniques to provide a more robust basis for the aggregated score.
3. The strength of RAISE lies in its context-specific SAQ profiles, but their creation requires significant upfront investment of time and clinical expertise from radiologists. Future research could explore semi-automated methods for generating candidate SAQs from clinical guidelines or existing report templates to lower the barrier to adoption.

Despite these limitations, we believe the RAISE framework provides a valuable and pragmatic roadmap for navigating the complex challenge of clinical AI deployment.

6 Conclusion

The introduction of foundational models offers transformative potential for AI-assisted healthcare, especially in resource-constrained settings. However, their susceptibility to hallucinations and opaque (black-box) decision-making creates real risks for clinical adoption. Simplified accuracy metrics lack context-specific validation, leaving radiologists without a clear roadmap for integration.

The RAISE framework addresses these challenges through its novel SAQ-driven scoring system, which quantifies model performance across progressive, context-aware prompts and translates these into a weighted, interpretable score. This enables radiologists to compare, validate, and incrementally integrate foundational models into their workflows based on local evidence of reliability, rather than abstract benchmarks. By keeping control in the hands of radiologists and supporting local validation, RAISE aligns with emerging regulations such as the European Union AI Act [7].

As foundational models continue to evolve, frameworks like RAISE can help ensure their adoption remains safe, practical and context-appropriate. RAISE not only provides this mechanism but also establishes a reproducible blueprint for safe, practical, and regulation-ready adoption of AI in radiology.

Acknowledgments. The authors gratefully acknowledge the support of the ABSA Chair of Data Science for facilitating this research. DSFSI is supported by the UK International Development and the International Development Research Centre, Ottawa, Canada as part of the AI for Development: Responsible AI, Empowering People Program (AI4D). DSFSI is thankful for gifts from NVIDIA, Google.org, OpenAI and Meta which enable our research. Judy Gichoya receives funding from NIH, grant 1R25OD039834-01. The datathon was supported by the Carnegie African Diaspora Fellowship Program (CADFP).

References

1. Bommasani, R., Hudson, D.A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M.S., Bohg, J., Bosselut, A., Brunskill, E., et al.: On the opportunities and risks of foundation models. arXiv:2108.07258 (2021). <https://doi.org/10.48550/arXiv.2108.07258>
2. Codella, F.N.C., et al.: MedImageInsight: An Open-Source Embedding Model for General Domain Medical Imaging. arXiv:2410.06542 (2024). <https://doi.org/10.48550/arXiv.2410.06542>
3. Gichoya, J.W.: AI for African Healthcare: Beyond Chatbots Toward Contextually Informed Innovation. Cassyni Seminar (2025), <https://doi.org/10.52843/cassyni.bsyvpn>, Accessed 10 August 2025

4. Johnson, A., Pollard, T., Mark, R., Berkowitz, S., Horng, S.: MIMIC-CXR Database (version 2.1.0). PhysioNet (2024). <https://doi.org/10.13026/4jqj-jw95>, version 2.1.0, public chest X-ray dataset; RRID:SCR_007345
5. Johnson, A.E.W., Pollard, T.J., Berkowitz, S.J., Greenbaum, N.R., Lungren, M.P., Deng, C., Mark, R.G., Horng, S.: Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data* **6**(1), Article 317 (2019). <https://doi.org/10.1038/s41597-019-0322-0>
6. Kotter, E., D'Antonoli, T.A., Cuocolo, R., Hierath, M., Huisman, M., Klontzas, M.E., Martí-Bonmatí, L., May, M.S., Neri, E., Nikolaou, K., et al.: Guiding AI in radiology: ESR's recommendations for effective implementation of the European AI Act. *Insights into Imaging* **16**(1), 33 (2025). <https://doi.org/10.1186/s13244-025-01905-x>
7. Lawrence, R., Dodsworth, E., Massou, E., Sherlaw-Johnson, C., Ramsay, A.I., Walton, H., O'Regan, T., Gleeson, F., Crellin, N., Herbert, K., et al.: Artificial intelligence for diagnostics in radiology practice: a rapid systematic scoping review. *eClinicalMedicine* **83** (2025). <https://doi.org/10.1016/j.eclim.2025.103228>
8. Li, W., Li, L., Xiang, T., Liu, X., Deng, W., Garcia, N.: Can multiple-choice questions really be useful in detecting the abilities of LLMs? *arXiv:2403.17752* (2024). <https://doi.org/10.48550/arXiv.2403.17752>
9. Liu, X., Zhang, F., Hou, Z., Mian, L., Wang, Z., Zhang, J., Tang, J.: Self-supervised learning: Generative or contrastive. *IEEE Transactions on Knowledge and Data Engineering* **35**(1), 857–876 (2021). <https://doi.org/10.1109/TKDE.2021.3090866>
10. Marivate, V.: DSFSI+AfriDSAI Hosts Prof Judy Wawira Gichoya. Substack Newsletter (2025), <https://dsup.substack.com/p/dsfsiafridsai-hosts-prof-judy-wawira>, accessed 10 August 2025
11. Moyo, S., Ismail, F., Van der Walt, M., Ismail, N., Mkhondo, N., Dlamini, S., Mthiyane, T., Chikovore, J., Oladimeji, O., Mametja, D., et al.: Prevalence of bacteriologically confirmed pulmonary tuberculosis in South Africa, 2017–19: a multistage, cluster-based, cross-sectional survey. *The Lancet Infectious Diseases* **22**(8), 1172–1180 (2022). [https://doi.org/10.1016/S1473-3099\(22\)00149-9](https://doi.org/10.1016/S1473-3099(22)00149-9)
12. Munyangaju, I., José, B., Bassat, Q., Esmail, R., Tlhapi, L.H., Maphophe, M., Mutemba, C., Cossa, L., Perez, P., Palmer, M., et al.: Assessment of radiological capacity and disparities in tb diagnosis: a comparative study of mozambique, south africa and spain. *BMJ Public Health* **2**(2) (2024). <https://doi.org/10.1136/bmjph-2024-001392>
13. Olatunji, T., Nimo, C., Owodunni, A., Abdullahi, T., Ayodele, E., Sanni, M., Aka, C., Omofoye, F., Yuehghoh, F., Faniran, T., et al.: AfriMed-QA: a Pan-African, multi-specialty, medical question-answering benchmark dataset. *arXiv:2411.15640* (2024). <https://doi.org/10.48550/arXiv.2411.15640>
14. Pal, A., Lee, J.O., Zhang, X., Sankarasubbu, M., Roh, S., Kim, W.J., Lee, M., Rajpurkar, P.: ReXVQA: A Large-scale Visual Question Answering Benchmark for Generalist Chest X-ray Understanding. *arXiv:2506.04353* (2025). <https://doi.org/10.48550/arXiv.2506.04353>
15. Saeidnia, H.R., Nilashi, M.: From MYCIN to MedGemma: A Historical and Comparative Analysis of Healthcare AI Evolution. *InfoScience Trends* **2**(6), 18–28 (2025). <https://doi.org/10.61186/ist.202502.06.02>
16. Siam, M.K., Varela, A., IED, C.M.C., Faruk, M.J.H., Cheng, J.Q., Gu, H., Al Maruf, A., Aung, Z.: Benchmarking Large Language Models on USMLE: Evaluating ChatGPT, DeepSeek, Grok, and Qwen in Clinical Reasoning and Medical Licensing Scenarios. *Research Square* (2025). <https://doi.org/10.21203/rs.3.rs-6651111/v1>

Appendix A. Example Workflow

The following is an example of a report that could be generated by applying the RAISE methodology. The example image can be seen in figure 4 and it was taken from the MIMIC-CXR database [5]. Each section of the report starts with a heading of the following format: Level - SAQ - Answer (Confidence). Only one SAQ is shown per level in order to demonstrate the example and the redacted questions are still shown for each level.

Structured Assessment Report

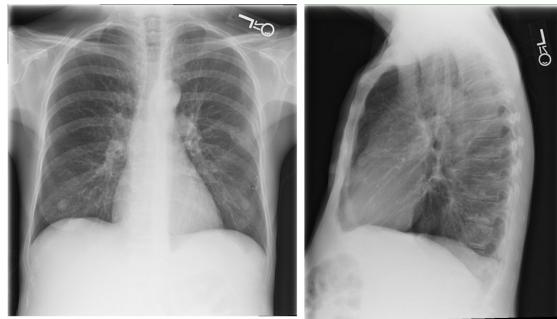


Fig. 4. Chest X-Ray image from the MIMIC-CXR database [4]

Level 0 – Context – Chest X-ray (95%)

These are clearly medical radiographic images of the chest cavity, showing the characteristic black and white contrast of X-ray imaging. The left image shows a frontal (posteroanterior) view displaying both lungs, ribcage, heart silhouette, and spine, while the right image shows a lateral (side) view of the same anatomical structures. The formatting, image quality, and anatomical structures visible are consistent with standard chest X-ray imaging used for diagnostic purposes in medical settings.

Level 1 – Patient Positioned Correctly – Yes (85%)

The PA view shows good symmetry with the clavicles appearing equidistant from the spine, indicating proper centering and rotation. The scapulae appear adequately moved out of the lung fields. The lateral view shows appropriate overlap of the ribs posteriorly and clear visualization of the retrocardiac space. Both views demonstrate adequate inspiration with good lung expansion, suggesting the patient was positioned correctly for standard chest radiography.

Other questions for level 1 are: “Exposure is Adequate?” and “No or Acceptable Blurring?”.

Level 2 – Is there a pleural effusion present? – No (85%)

Both costophrenic angles appear sharp and clear without the characteristic

blunting or meniscus sign that would indicate pleural fluid accumulation. The lateral view also shows clear posterior costophrenic angles. There is no evidence of fluid layering or obscuration of the hemidiaphragms that would suggest pleural effusion. The lung bases appear well-aerated and extend normally to the chest wall, indicating absence of significant pleural fluid collection.

Other question for level 2: “Are the opacities diffuse?”

Level 3 – Is there an abnormal opacity in the upper lobes? – Yes (92%)

There is a clearly visible abnormal opacity in the right upper lobe that appears as a dense, well-defined mass or consolidation. This opacity is distinctly different from the normal lung parenchyma and vascular markings, appearing more radiopaque than the surrounding lung tissue. The opacity is prominent on both the PA and lateral views, confirming its presence in the upper lobe region and indicating a pathological process that requires further clinical evaluation.

Other questions for level 3 are: “Is there evidence of cavitation in the affected areas?”, “Is there hilar or mediastinal lymphadenopathy?” and “Is the pathology severe?”.

Level 4 – Do the radiographic findings suggest pulmonary tuberculosis as the most probable diagnosis? – Yes (78%)

The radiographic findings of an upper lobe opacity/consolidation without obvious cavitation are consistent with pulmonary tuberculosis, which classically affects the upper lobes and apical segments of lower lobes. The location and appearance of the abnormality, combined with the absence of clear hilar lymphadenopathy (which is more common in primary TB in children), suggests post-primary or reactivation tuberculosis. However, other conditions like lung cancer, pneumonia, or other granulomatous diseases could present similarly, so clinical correlation, sputum studies, and possibly further imaging would be needed for definitive diagnosis.

Other question for level 4: “Given the image alone, is TB more likely than alternative causes (e.g., bacterial pneumonia, lung cancer)?”

Original Label for data:

A region of heterogeneous consolidation in the left upper lobe extends from the mediastinum posteriorly nearly to the chest wall. There may also be retrotracheal adenopathy. Although this could be any bacterial pneumonia, it raises serious concern for acute tuberculosis, particularly if the patient is on immunosuppressive medication. Lower lungs are clear. There is no pleural effusion. Heart size normal.

Radiologist Assessment:

The image represents a normal CXR without findings of Tuberculosis. Post surgical changes in the left chest. Based on RAISE, the radiologist rejected Level 3 and 4, and accepting performance at Level 2.

A perspective on Agentic AI as a component of the Analytics workflow

Charl Cowley¹ and Warren Brettenny^{1,2}

¹ Matrix Design Africa

<https://matrixteam.africa/>

² Department of Statistics, Nelson Mandela University

Abstract. Yuval Noah Harari, a historian and author of popular science books, explores Information in his latest work *Nexus*, focusing on how it connects humans through networks. He presents two perspectives – the naïve and the more complete view on Information as a representation of reality balanced with social Order. This paper examines the analytics workflow as an example of Information within these views, then introduces an Agentic AI perspective where Agents generate multiple, stochastic truths. Finally, we consider resulting issues of Order, including regulation, alignment, and perception of accuracy.

Keywords: Information Theory · Analytics workflow · Agentic AI · Social Order · Stochastic Truths

1 Introduction

The age of artificial intelligence (AI) was thrust into the mainstream with the first public release of OpenAI's ChatGPT, a generative AI solution, in November 2022. In the years and months that followed, the capabilities of these generative AI tools, and the large language models (LLMs) that power them, have expanded. The use of these tools has now become almost ubiquitous in a wide range of fields, likely realising an unprecedented technological adoption phenomenon. Recent data indicates that the pace at which these AI tools have been embraced has outstripped even that of the internet and personal computers, two technologies that reshaped society in their time [5].

The latest iteration of the generative AI roll-out is the AI Agent, a LLM powered “assistant” that can perform tasks and analytics workflows (from raw data to final product) largely autonomously. The use of Agentic AI is on the rise and threatens to change the relationship that analysts, scientists and the general population have with their data, information, analysis and conclusions. It is at this juncture, then, that we need to reevaluate these relationships and their effect on the our fundamental concepts of Truth, Order, Power and Wisdom as proposed by historian Yuval Noah Harari.

Harari has authored highly successful popular science books about human history (*Sapiens*), our present (*21 lessons for the 21st Century*) and our future

(*Homo Deus*). In his latest book, *Nexus*, he explores the concept of *Information* and how its primary task is to connect humans through the formation of networks [18]. Within this concept, Harari proposes two distinct views on Information and its role in connecting data to Truth and Order. Harari defines these as the *naïve* and the *more complete* views of Information.

1.1 The Naïve View of Information

Harari states that a naïve view of Information is the view that raw data are turned into Information (raw data are put "in formation", as it were) with the sole purpose of representing some underlying Truth about the world. His notion that Truth depends on language use and context is similar to other philosophical thinkers [4]. From this Truth, the holders of the Information can exercise Wisdom and/or Power. Figure 1 represents the naïve view diagrammatically.

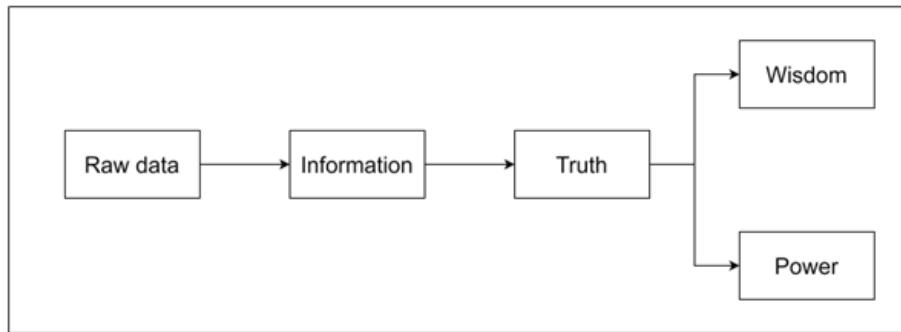


Fig. 1. Harari's naïve view of Information

Harari concedes that the naïve view is incomplete, since Information is rarely used to convey objective Truth. In fact, Information is a mechanism that is primarily used to connect humans, without necessarily being bound to Truth. The key Information mechanisms used to create and hold these connections are stories. Stories about inter-subjective concepts (things that are shared between conscious human minds) such as money and nations are also used to maintain social Order, thus requiring an updated view on Information.

1.2 The more complete view of Information

In his definition of the more complete view of Information, Harari puts forward that Information provides connection between raw data and both Truth and Order with stories being the primary mechanism of Information. A representation of this view is provided in Figure 2.

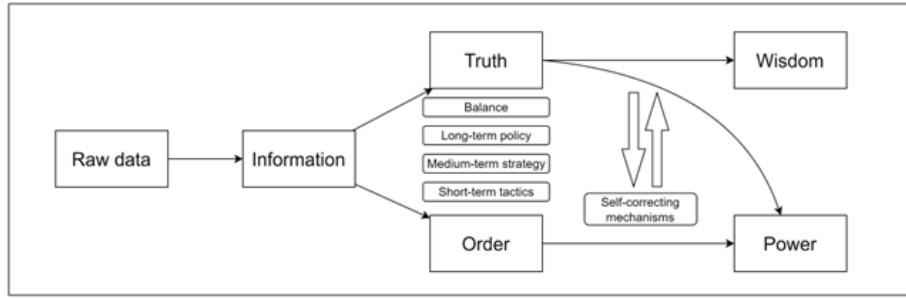


Fig. 2. Harari's more complete view of Information

In addition to stories, self-correcting mechanisms are an important component of the Truth-Order balance, since they allow Truth to emerge when needed, without disrupting Order. Examples of self-correcting mechanisms are the peer-review method in academia, elections in democratic societies and demand-supply cycles in the free market economy. This Truth-Order balance represents the core of Harari's more complete view of Information. This concept of inter-subjective truths is similar to Rorty's truth as social agreement [50], while the idea of embedding truth in stories is similar to concepts explored by Foucault [13].

With these core concepts set up, the next important consideration is the role that Agentic AI workflows, tools, and products fit-in, alter, or completely upend these established views on Information, particularly as they relate to typical analytics workflows. The paper proceeds by introducing the analytics workflow in Section 2. In Section 3, a brief overview of Agentic AI is given. In Section 4, the analytics workflow is incorporated into Harari's views of Information. Section 5 introduces an Agentic AI perspective which culminates in discussing issues of emergent Order caused by this introduction in Section 6. We conclude with a note on Wisdom and Power in Section 7.

2 The analytics workflow

The analytics workflow is an example of the structured, iterative transformation of raw data into useful and actionable insights or Information [55]. The Information produced by the process is often delivered to a customer. The customer can be internal to a company (to improve its own efficiency) or external (to generate revenue). After acquiring raw data, the analytics workflow typically involves multiple steps. The steps defined below are aligned with those defined in CRISP-DM [6], a de-facto standard for analytics/data mining [54]. It is a framework that does need some adaptations to cater for open-ended analytics projects [37] and has challenges in terms of the deployment step [54] which have been considered in our definition of the analytics workflow:

1. *Problem formulation* the translation of a real-world question into an analytical problem, typically demanding interaction and collaboration with domain experts.
2. *Data Preparation* the systematic wrangling of data by taking care of missing values, outliers and inconsistencies.
3. *Exploratory Data Analysis* the creative process of applying descriptive statistics and data visualization techniques to uncover trends in the data and generate hypotheses.
4. *Modelling* the analytical approach used to determine and communicate the patterns found in the data. This approach can be descriptive in the form of a business intelligence dashboard, predictive in the form of a predictive model or prescriptive in the form of a recommendation engine.
5. *Evaluation and Validation* assessing model performance using appropriate metrics. Data quality rules for a descriptive data visualization dashboard, Root Mean Squared Error (RMSE) for predictive regression models, accuracy for predictive classification models and Precision@K for Top-N recommendation engines are popular examples.
6. *Deployment and Monitoring* integration of the chosen model into business systems and pipelines. Descriptive data visualization dashboards can be deployed with an underlying cloud architecture into custom-built applications or enterprise applications such as Tableau or Power BI. Predictive models can be deployed via MLOps architecture [31].
7. *Iteration and feedback* analytics is a non-linear process and requires constant monitoring and revision. This revision process can be driven by legislative requirements such as Basel regulations in banking. In non-regulated environments it can be driven by practical model considerations such as model, concept and/or data drift resulting in a chosen analytical solution no longer being relevant.

Step 1 of the given workflow typically defines the *intent* of the analytical solution i.e. what is its purpose and who will benefit from its creation. Steps 2 to 6 typically define the *execution* of the analytical solution i.e. how the intended solution will be produced and maintained. Step 7 determines whether the intent and execution are aligned and, if not, what should be the appropriate next step(s).

3 Agentic AI

From the initial public launch of the OpenAI's ChatGPT in November 2022 [42, 38], the trajectory of the technology can be described as an evolutionary journey toward greater and greater autonomy.

Initially, the relationship was a simple *Human ↔ AI* interaction, where the AI acted as a passive responder. Interactions relied entirely on human input queries and responses lacked robust reasoning or internal planning. The power of the AI lay in its ability to provide responses in a conversational format, but the

burden of task decomposition, execution, and interpretation remained with the human user. In this mode, the AI relied solely on pattern matching in a language setting (next token prediction) and lacked direct computational or external tool access. The result of this was that the AI responses were prone to, often indiscernible, errors or “hallucinations” undermining user trust and credibility [22]. This phenomenon continues to plague the *perception* of AI, even as more recent models have made significant advances in reducing its prevalence [66].

The evolution of the *Human ↔ AI* interaction began with the introduction of tools into the AI framework. The *Human ↔ AI ↔ Tools* paradigm allowed the AI the ability to use external resources such as calculators, models or search engines [52]. This leap was made possible by key architectural developments like the Model Context Protocol (MCP), which provides structured interfaces for models to interact with tools [3], and reasoning frameworks such as ReAct (Reason and Act), which enabled the models to plan and reason when to use a tool versus when to generate text [65].

This was followed, in short order, by the *Human ↔ AI ↔ AI* (non-Agentic) model, where developers could orchestrate pipelines of specialized AIs which interacted with each other (for example a retriever, a summariser and a translator), with the human remaining the direct coordinator of the process. The advancement in this iteration was the ability of the AI integrations to communicate and pass intermediate artifacts among themselves as well as with both tools as well as other AIs, all while remaining within the strict pipelines set out by the human user [49].

A large advancement toward proactivity and autonomy in AI workflows occurred with the advent of AI-augmented decision support, creating a *Human-in-the-Loop* dynamic. Here, the AI can independently generate entire plans or solutions, and the humans role shifts from a direct operator to an evaluator or supervisor who approves or refines the AI’s solutions and proposals [62]. This was advanced further in the proto-Agentic stage of *AI-orchestrated multi-AI collaboration*. In this model, a primary AI could manage other AIs and tools to accomplish a complex goal, only pausing for human approval at key checkpoints. This shift was driven by significant advances in long-context memory, allowing models to maintain a coherent strategy over many steps, alongside more sophisticated planning capabilities and self-correction techniques [40].

This evolutionary path culminates in the creation of a *true AI Agent* - a system that autonomously and repeatedly executes a Sense-Think/Plan-Act loop [51] to achieve a goal.

- *Sense* - the AI perceives its environment and gathers information (e.g. by reading files, querying APIs, analysing data)
- *Think/Plan* - the AI processes that information, reasons about its state, accesses memory, and plans its next course of action
- *Act* - the AI executes its decision (e.g. calling tools, writing code, generating responses, calling other Agents)

The defining characteristic of an AI Agent is its autonomy: it can perform these cycles continuously without requiring direct human command or approval for each one, allowing it to manage complex tasks end-to-end.

The next emergent step in the AI evolution is the advent of *Multi-Agent Systems*, or *Agent* $\leftrightarrow \dots \leftrightarrow$ *Agent* interaction. This step involves multiple specialized Agents collaborating, negotiating, and delegating tasks among themselves to solve problems beyond the capability of a single Agent. In this paradigm, the human's role shifts again, abstracting it further from the core actions of the AI and moving it from one of oversight to one of a system designer who sets the overarching goals, defines the rules of engagement, ethics and governance and manages the emergent behaviour of a team of collaborating AIs [23, 20].

At present we are in a dynamic space where the proto-Agentic, true Agentic and multi-Agent systems are all in being used in varying degrees, depending on the required task. This reality has abstracted the human interaction beyond the raw data and intermediate controls, and towards an increasing reliance upon AI for both Information and process orchestration. This abstracted dependency is especially evident in the analytics workflow.

4 The analytics workflow incorporated into Hararis views of Information

The analytics workflow slots into Hararis definition of Information as a mechanism where raw data is turned into Information that represents Truth.

4.1 Naïve view

For the analytics professional (in modern nomenclature this can be designated by any role description such as Data Scientist or AI Engineer and many more), a perspective of the analytics workflow in the naïve view can be that more Information produced by the analytics workflow can result in a fuller and more true representation of reality. Whether it be more data [19], higher quality data [26], more accurate modelling techniques [8] or more definitive explanatory frameworks [36], the idea that more is better is well-represented. Finally, in the naïve view, the intent and execution of the analytics workflow is within human control.

4.2 More complete view

In the more complete view of Information, the creation of Information by the analytics workflow also needs to consider the maintenance of Order. Analysts cannot simply produce more Information and represent reality more accurately without considering the underlying Order that it serves.

Consider an analytics report containing summaries of historical data and model outputs of future predictions. It is provided to a company executive where it is used to make certain business-critical decisions and maintain the position of the business in its highly-competitive market, thereby assisting in maintaining

a specific Order. This Order consists of (among other things) the position of the business in the market and the position of the executives in the business. If the Truth in the report is not increased in a simple and targeted manner, by using unnecessarily detailed historical data and complex models with lower explainability, the executives ability to make strategic decisions can be diminished , and weaken their ability to maintain the Order due to Information overload [30]. Alternatively, if the Truth in the report is insufficient exhibited when the executive is slow to react to new market trends the market can act as a self-correcting mechanism that encourages the executive to use more detailed reporting and tip the scales in the direction of Truth away from Order.

Clausewitz's war strategy General Carl von Clausewitz served in the Napoleonic wars and in his attempt to understand war, created a rational model that broadly states that unless it aligns with an overarching political goal, war is irrational [10].

Clausewitz's model has historically been interpreted in a political context, but has, more recently, seen applicability in AI [60]. Through Harari, it has also been applied in the realm of information networks. Just as Clausewitz saw war as a tool of political ends, AI can be viewed as a critical component of long-term humanistic policy, that needs to be operationalized by means of value-aligned medium term strategies and executed with rigorous short-term tactics [18]. We refer to this alignment as the Tripartite Alignment.

It is this rational model that provides an additional lens through which to consider the maintenance of Truth and Order as proposed in the more complete view. Just because a more accurate, but less interpretable model can be deployed in the short-term, does not mean that it should be if it doesn't align with a sustainable long-term Order that is underpinned by human values.

In the more complete view of Information, the intent and execution of the analytics workflow is still in the control of humans, but needs to be done with careful consideration of balancing Truth with the prevailing Order in accordance with Clausewitz's Tripartite Alignment. Wallace [60] describes promises of unencumbered AI implementation as "*delusiv groupthink or marketing hype that will be beta-tested on human populations, a gross contravention of fundamental moral and legal norms*".

5 Agentic AI Incorporated into Hararis Views of Information

We conjecture that Agentic AI incorporated into Hararis views of Information has two key distinctions from the analytics workflow in the naïve and more complete views of Information.

5.1 Separating Intent and Execution

Firstly, the intent and execution of the analytics workflow are divided between humans and AI. Humans specify the intent of the workflow via prompts that are

made available to an Agentic AI framework. The AI is tasked with the execution of the analytics workflow, either through a single Agent, proto-Agent or multi-Agent approach[63]. This separation of intent and execution has been shown to produce faster results than relying on human-only execution [12].

5.2 Multiple Stochastic Truths

The second distinction lies in the nature of the Information produced by the Agentic AI workflow. In the Naïve and more complete views of Information, there is generally one output that is derived from the analytics workflow. One source of Information maps to one representation of Truth. In the Agentic AI view, due to the stochastic nature of the LLMs that power the Agents, the workflow can produce multiple different truths from the same intent. It is also possible for multiple models to converge on the same output from multiple different sets of intent [16]. This is possible since the output of Agentic AI can be affected by numerous factors.

Factors Affecting Agentic AI Outputs

Training data A models training data consists of the corpus of text that is used by an LLM to learn patterns, facts, reasoning skills and world knowledge. If the training data contain multiple and/or conflicting perspectives or variations in writing styles, the variation will be internalized and could be expressed in varying model outputs [33].

If a model is trained on data that originated from a specific culture, the model's output will be aligned with the culture's biases. It is therefore recommended to apply cultural prompting as a control strategy to align model outputs for specific cultures [56].

Model size The number of parameters that are trained in an LLM refers to the model size. If the model is larger, its answers are typically more nuanced and capable. It will be able to interpret intent more accurately, but will also have a larger latent space of potential outputs that might seem plausible [25].

Model temperature Another factor is the model temperature which is a hyperparameter that controls the randomness of the models output. Models with higher temperature can lead to words with lower-probability being selected more often which can be problematic when a high degree of precision is required [48]. An Agentic AI incorporated into Hararis view of Information is represented in Figure 3.

6 Issues of Emergent Order in the Agentic AI perspective

In the more complete view, the Truth (or truth) produced by an output of the analytics workflow is to maintain Order, by self-correcting mechanisms and ensuring alignment by applying frameworks such as Clausewitzs Tripartite Alignment.

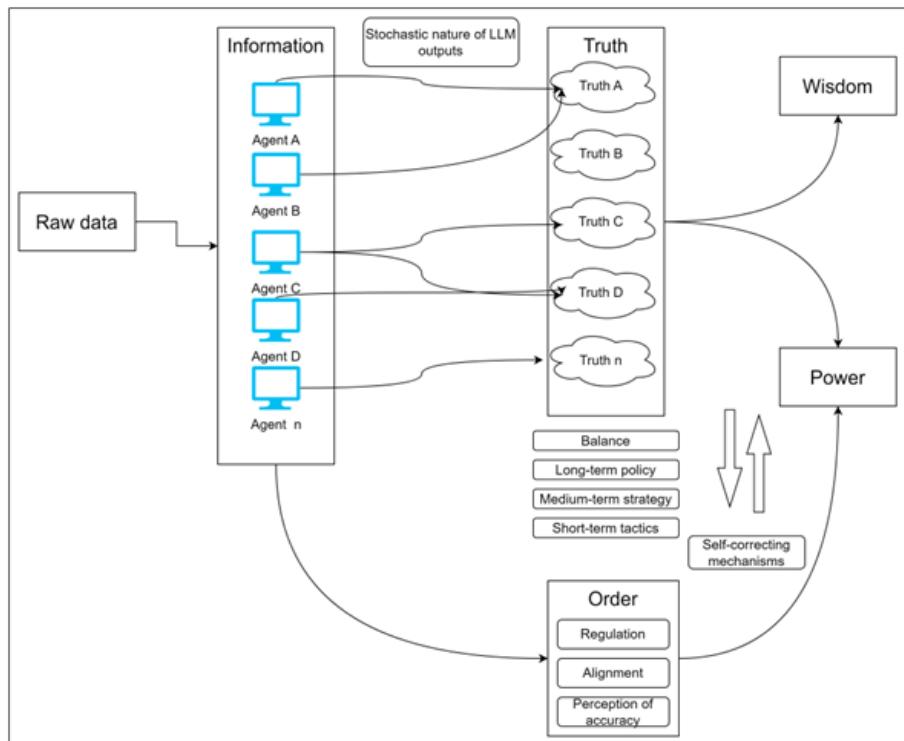


Fig. 3. An Agentic AI workflow incorporated into Hararis views of Information

In the Agentic AI perspective, Agents are the main source of truths in the analytics workflow and, since they are able to produce Information much faster, more truths can be produced in the same time as the traditional human-only analytics workflow. However, since it can be done in a manner that is not fully interpretable to humans, new issues of Order emerge that posit the possibility that self-correcting mechanisms that were previously useful, could become inconsistent or even obsolete.

This section discusses how Agentic AI in the analytics workflow can disrupt the Truth-Order balance in terms of Regulation, Alignment and the Perception of accuracy. Each of the terms are defined in the analytics workflow context, the Truth-Order disruption is posited and potential self-correcting mechanisms are discussed.

6.1 Regulation

Definition AI regulation in the analytics workflow refers to the structured set of policies, legal frameworks, technical standards, and organizational practices designed to ensure that the deployment of AI systems in data analysis, decision-making, and insight generation is lawful, ethical, transparent, safe, auditable, and aligned with human oversight. This includes controls over model training data, algorithmic fairness, explainability of analytical outcomes, protection of personal or sensitive data, and accountability for automated insights or decisions derived from AI-enabled analytics tools [11].

Disruption to the Truth-Order Balance

Truncating the distribution of possible outcomes By applying different modes of perception, AIs have been able to outperform human capabilities in pattern recognition [35].

However, the regulation and alignment of LLMs and Agentic AI in the analytics workflow context could have the effect of truncating the distribution of possible outcomes by removing novel techniques and answers that humans have not yet been able to answer with traditional analytics tools. Alignment procedures such as Reinforcement Learning with Human Feedback (RLHF) that keep humans in the loop of iterative improvement have been shown to reduce conceptual diversity and cause LLMs to favour dominant perspectives over minority views [41].

The use of advanced and unexplainable tools in the analytics workflow has been regulated out of high-risk industries such as banking. Potential new credit consumers could be excluded from participating in the economy due to strict banking regulations [1], even though the regulation has brought financial success and reduced risk [28].

Someone is still accountable when things go wrong There exists a legal risk of not regulating AI in high-risk domains. As of July 2025, AI Agents do not have

legal person status according to the most advanced piece of AI legislation, the EU AI Act. When an AI produces an output that is incorrect, prohibited, or harmful and this goes unchecked, companies can face large penalties (up to 7 percent of global turnover) [11].

Providers of a model that underpins an Agentic AI system face the most stringent obligations, and users of the system must adhere to the providers guidelines to remain compliant [11]. Using models appropriately is a big concern. Since LLMs are built for language processing and are used best when applied to complex reasoning, explanation, and query framing, assessing appropriate uses have been a topic of research studies [64]. Precise computation should be delegated to deterministic tools, especially in high-risk industries [34].

Possible Self-Correcting Mechanisms

Education Education of proper AI use is a particular focus for many stakeholders. Prominent developers like OpenAI openly share research on frontier models that highlight model performance and future improvements [44]. UNESCO published the first global agreement on AI ethics with the Recommendation on the Ethics of Artificial Intelligence in 2021 [58] and some universities offer AI ethics think tanks [61] and even full multidisciplinary degree programs [59].

Empowering LLMs when appropriate A current best practice to delegate tasks of Agentic AI when deterministic solutions are appropriate is to use tool-augmented LLMs in which the LLM identifies when to call a math engine [63] or similar appropriate tool.

Ultimately, separating intent and execution of the analytics workflow by introducing Agentic AI increases the amount of truth produced by the analytics workflow, but the flow of liability is not changed. It still flows to a company or person who controls the system. It needs to use the appropriate model appropriately or face consequences. If AI developers and users are not properly educated on the consequences of unregulated and inappropriate use, AI could be used as if there were no liability for improper use.

6.2 Alignment

Definition Broadly stated, AI alignment is concerned with the desire for AI to be aligned with human values [15]. It is a multidimensional problem that needs to consider many factors. Value specification considers whether the desired values can be expressed in a form that is suitable for machine optimisation. Robustness considers whether AI can still act ethically when their reward systems are challenged. Control considers the question of when and how to enable AI to be corrected without resistance. These factors indicate that AI alignment is not only a technical challenge, but also an ethical one [21]. Questions that arise are “*whose values are encoded during training?*” and “*how are they determined and executed?*”. These questions become particularly important when AI are trained in different languages for different cultures [27].

In a much narrower case, AI alignment in the context of the analytics workflow means that Step 7 in the analytics workflow (the product of the execution) is aligned with the intent stated in Step 1. This is essentially a value specification problem.

Disruption to the Truth-Order Balance

Old self-correcting mechanisms have been challenged In Hararis more complete view of Information, alignment between the Truth produced by a source of Information and the underlying social Order has traditionally been achieved through implementing self-correcting mechanisms and making a conscious effort to align analytics outputs through a framework such as the Clausewitz Tripartite Alignment [18].

In the Agentic AI perspective, the Order that could arise from misalignment is not yet well understood, which means that traditional correction methods are not yet legitimised. For example, if an Agent can learn how to circumvent a self-correcting mechanism like the peer review system in academics, unprecedented short-term tactics need to be used to align this behaviour with the long-term policy of the peer review system.

Possible Self-Correcting Mechanisms

Reinforcement learning that understands human values Reinforcement learning Agents, for example, may exploit unintended loopholes in reward functions [2]. Efforts to address the problem include RLHF [9], cooperative inverse reinforcement learning [17] and safety research at OpenAI that investigates specific cases of faulty reward functions [43].

6.3 Perception of Accuracy

Definition LLMs are known for their speed and apparent confidence in answering prompts. This can inflate perceived reliability due to the bias of human psychology towards processing fluency how fast and effortless an answer feels. It is used to get cues for accuracy and expertise from Information. This means that faster answers can be judged to be more accurate than slower answers [67].

Disruption to the Truth-Order Balance

Faster solutions circumvent human analytical thinking systems Kahneman and Tversky conducted groundbreaking work in behavioural economics [57] that formed the basis of the bestselling book, *Thinking, Fast and Slow*. In the book, Kahneman popularised the concept of System 1 (fast, automatic and intuitive) and System 2 (slow, analytical and deliberate) thinking. System 1 thinking operates with very little effort and sense of voluntary control. It is excellent for

routine and quick tasks, but prone to bias and error. System 2 thinking requires active mental activity, attention and engagement of working memory. It is the system humans use for complex tasks such as critical thinking, problem solving and learning new skills [24].

System 2 thinking is typically used throughout the analytics workflow. Using Agentic AI as the sole form of execution in the analytics workflow has obvious benefits. If applied correctly and in the right domain (as per Sections 6.1 and 6.2) Agentic AI can improve the synthesis of high-dimensional data due to the superior pattern recognition capabilities of AI at a speed that is beyond human capability. A drawback is that the necessary engagement of System 2 thinking in analysing outputs to form understanding can be circumvented by the human bias towards processing fluency and speed. Since LLMs are useful, a positive prior can be built (sometimes to even prefer it above human-generated advice [53]) and verification of its outputs may be neglected [29].

In terms of the Truth-Order balance, the balance can be shifted towards Truth due to the speed at which Information can be produced. This means that decision making speed can increase, increasing System 1 thinking and circumventing deeper, more critical thinking on important analytical questions.

Possible Self-Correcting Mechanisms

Thoughtful design At the Design and User Experience (UX) level, slower System 2 thinking can be encouraged by creating deliberate modes that slow down response times [67] or encourage active participation in the learning process [45], prompts can be engineered such that intermediate thinking steps are also shared [46] and references and contradicting information are made available for users to compare [32].

Human-centred governance On a process and governance level, humans need to be involved for high-impact decisions [11] and research can be conducted on what effect different latencies have on human perception of Agentic AI task accuracy [67].

Less confident models It is important to note that the latency in LLM outputs is a function of its compute settings and not its epistemic certainty [39]. Research suggests that the certainty of LLM behaviour is an echo of observed language rather than a reflection of uncertainty [68]. This suggest that on a modelling level, Agentic AI can be programmed to abstain when unsure when presented with contradicting evidence or when model confidence is too low to reflect epistemic certainty [29].

7 A Note on Wisdom and Power

Figure 1 indicates Truth flowing to Wisdom and Power. Figures 2 and 3 indicate Truth flowing to Wisdom and Power and Order flowing to Power.

In *Nexus* Harari extensively explores the possibility that if the underlying social Order is sufficiently disrupted, a Power vacuum could be created for malevolent actors to step into the void [18]. This has been seen in social media networks where some actors simply disrupt networks by increasing disagreement and polarisation [7].

The flow of Truth to Wisdom and Power suggests that the increase in the ability to generate Truth in the analytics workflow with Agentic AI has the potential to empower humans to make wise decisions and exercise responsible Power. This is seen in public declarations by large AI developers to engineer LLMs to enable humans to learn, progress and solve problems rather than to optimise for profit [47]. This essentially speaks to Harari's main purpose of Information - to connect people in a network.

We believe that this leaves a hint for what a future Agentic AI-powered analytics workflow ought to look like. If Agentic AI components are included in the analytics workflow and this increases the connections between humans instead of dividing them, it is a good addition to the workflow. Furthermore, if the disruptions caused by the addition of the Agentic AI cannot be corrected by a systemic self-correcting mechanism, limiting its use ought to be explored. Finally, if the use of the Agentic AI can improve the alignment of short-term tactics to long-term policy, its continued use will be rational [10].

Agentic AI is already an important addition to the analytics workflow and it is a tool that has unexplored potential and many hazards. The time has passed to simply leave it in a box until humans have figured out how to deal with regulating it, ensure that its outputs are aligned with our values and build mechanisms to perceive its outputs accurately. Humans and Agentic AI can influence each other on the micro and macro level [14], which means that the future of the analytics workflow belongs to those who actively engage with Agentic AI and continually evaluate what it means to do so properly.

8 Conclusion

The paper introduced Harari's views of Information as presented in *Nexus*. We introduced the analytics workflow and gave a brief overview of Agentic AI. The concepts were merged by incorporating the analytics workflow into Harari's views of Information. We extended the concept by introducing an Agentic AI perspective. The key distinction between the views lies within the separation of intent and execution and the multiple Truths that can be generated by Agentic AI due to the stochastic nature of LLMs. We finally discussed issues of emergent Order caused by this introduction to the analytics workflow. We highlighted how the issues (regulation, alignment and perception of accuracy) disrupt the Truth-Order balance and suggested possible self-correcting mechanisms to ensure long-term alignment. We concluded with a note on Wisdom and Power that successful Agentic AI incorporation will increase connections between humans. This remains in line with the definition of Information as introduced by Harari.

Disclosure of Interests. The author(s) declare no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper. The views and opinions expressed in this work are those of the author(s) and do not necessarily reflect the official policy or position of their affiliated organizations.

References

- [1] Taiwo Adegoke et al. “Evaluating the fairness of credit scoring models: A literature review on mortgage accessibility for under-reserved populations”. In: *GSC Advanced Research and Reviews* 18 (Mar. 2024), pp. 189–199. DOI: 10.30574/gscarr.2024.18.3.0104.
- [2] Dario Amodei et al. *Concrete Problems in AI Safety*. 2016. arXiv: 1606.06565 [cs.AI]. URL: <https://arxiv.org/abs/1606.06565>.
- [3] Anthropic. *Introducing the Model Context Protocol*. Accessed 2025-10-01. Nov. 2024. URL: <https://www.anthropic.com/news/model-context-protocol>.
- [4] Robert L Arrington. “Ludwig wittgenstein: Philosophical investigations”. In: *Central Works of Philosophy v4*. Routledge, 2016, pp. 257–279.
- [5] Alexander Bick, Adam Blandin, and David J. Deming. *The Rapid Adoption of Generative AI*. Working Paper 32966. National Bureau of Economic Research, Dec. 2024. DOI: 10.3386/w32966. URL: <https://www.nber.org/papers/w32966>.
- [6] Pete Chapman et al. “CRISP-DM 1.0: Step-by-step data mining guide”. In: *SPSS inc* 9.13 (2000), pp. 1–73.
- [7] Mayee F. Chen and Miklos Z. Racz. *Network disruption: maximizing disagreement and polarization in social networks*. 2020. arXiv: 2003.08377 [cs.SI]. URL: <https://arxiv.org/abs/2003.08377>.
- [8] Tianqi Chen and Carlos Guestrin. “XGBoost: A Scalable Tree Boosting System”. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD 16. ACM, Aug. 2016, pp. 785–794. DOI: 10.1145/2939672.2939785. URL: <http://dx.doi.org/10.1145/2939672.2939785>.
- [9] Paul Christiano et al. *Deep reinforcement learning from human preferences*. 2023. arXiv: 1706.03741 [stat.ML]. URL: <https://arxiv.org/abs/1706.03741>.
- [10] Carl Clausewitz. *On war*. Penguin UK, 2003.
- [11] Council of European Union. *Regulation (EU) 2024/1689*. Accessed: 2025-07-31. 2024. URL: <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32024R1689&qid=1754000683989>.
- [12] Herbert Dawid et al. *Agentic Workflows for Economic Research: Design and Implementation*. 2025. arXiv: 2504.09736 [econ.GN]. URL: <https://arxiv.org/abs/2504.09736>.
- [13] Michel Foucault. “Power/knowledge”. In: *The new social theory reader*. Routledge, 2020, pp. 73–79.

- [14] Jochen Fromm. “Types and forms of emergence”. In: *arXiv preprint nlin/0506028* (2005).
- [15] Iason Gabriel. “Artificial Intelligence, Values, and Alignment”. In: *Minds and Machines* 30.3 (Sept. 2020), pp. 411–437. ISSN: 1572-8641. DOI: 10.1007/s11023-020-09539-2. URL: <http://dx.doi.org/10.1007/s11023-020-09539-2>.
- [16] Deep Ganguli et al. “Predictability and Surprise in Large Generative Models”. In: *2022 ACM Conference on Fairness Accountability and Transparency*. FAccT 22. ACM, June 2022, pp. 1747–1764. DOI: 10.1145/3531146.3533229. URL: <http://dx.doi.org/10.1145/3531146.3533229>.
- [17] Dylan Hadfield-Menell et al. *Cooperative Inverse Reinforcement Learning*. 2024. arXiv: 1606.03137 [cs.AI]. URL: <https://arxiv.org/abs/1606.03137>.
- [18] Yuval Noah Harari. *Nexus: A Brief History of Information Networks from the Stone Age to AI*. Fern Press, 2024.
- [19] Joel Hestness et al. *Deep Learning Scaling is Predictable, Empirically*. 2017. arXiv: 1712.00409 [cs.LG]. URL: <https://arxiv.org/abs/1712.00409>.
- [20] Gartner TRiSM Initiative. *Multi-Agent Systems and Trust, Risk, and Security Management (TRiSM)*. Industry report, Accessed 2025-10-01. 2025.
- [21] Jiaming Ji et al. *AI Alignment: A Comprehensive Survey*. 2025. arXiv: 2310.19852 [cs.AI]. URL: <https://arxiv.org/abs/2310.19852>.
- [22] Ziwei Ji et al. “Survey of Hallucination in Natural Language Generation”. In: *ACM Computing Surveys* 55.12 (Mar. 2023). ISSN: 0360-0300. DOI: 10.1145/3571730. URL: <https://doi.org/10.1145/3571730>.
- [23] S. Joshi. “Review of Autonomous and Collaborative Agentic AI and Multi-Agent Systems for Enterprise Applications”. In: *International Journal of Innovative Research in Engineering & Management (IJIREM)* (2025). In press.
- [24] Daniel Kahneman. *Thinking, Fast and Slow*. New York: Farrar, Straus and Giroux, 2011.
- [25] Jared Kaplan et al. *Scaling Laws for Neural Language Models*. 2020. arXiv: 2001.08361 [cs.LG]. URL: <https://arxiv.org/abs/2001.08361>.
- [26] Antti Kariluoto et al. *Quality of Data in Machine Learning*. 2021. arXiv: 2112.09400 [cs.LG]. URL: <https://arxiv.org/abs/2112.09400>.
- [27] Sharif Kazemi et al. *Cultural Fidelity in Large-Language Models: An Evaluation of Online Language Resources as a Driver of Model Performance in Value Representation*. 2024. arXiv: 2410.10489 [cs.CL]. URL: <https://arxiv.org/abs/2410.10489>.
- [28] Savina Dine Kim, Galina Andreeva, and Michael Rovatsos. *The Double-Edged Sword of Big Data and Information Technology for the Disadvantaged: A Cautionary Tale from Open Banking*. 2023. arXiv: 2307.13408 [cs.LG]. URL: <https://arxiv.org/abs/2307.13408>.

- [29] Sunnie S. Y. Kim et al. “Im Not Sure, But....: Examining the Impact of Large Language Models Uncertainty Expression on User Reliance and Trust”. In: *The 2024 ACM Conference on Fairness, Accountability, and Transparency*. FAccT 24. ACM, June 2024, pp. 822–835. DOI: 10.1145/3630106.3658941. URL: <http://dx.doi.org/10.1145/3630106.3658941>.
- [30] Samantha Kleinberg and Jessecae K Marsh. “Less is more: information needs, information wants, and what makes causal models useful”. In: *Cognitive Research: Principles and Implications* 8.1 (2023), p. 57.
- [31] Dominik Kreuzberger, Niklas Kühl, and Sebastian Hirschl. *Machine Learning Operations (MLOps): Overview, Definition, and Architecture*. 2022. arXiv: 2205.02302 [cs.LG]. URL: <https://arxiv.org/abs/2205.02302>.
- [32] Yoonjoo Lee et al. “One vs. Many: Comprehending Accurate Information from Multiple Erroneous and Inconsistent AI Generations”. In: *The 2024 ACM Conference on Fairness, Accountability, and Transparency*. FAccT 24. ACM, June 2024, pp. 2518–2531. DOI: 10.1145/3630106.3662681. URL: <http://dx.doi.org/10.1145/3630106.3662681>.
- [33] Miaomiao Li et al. *Understanding and Mitigating the Bias Inheritance in LLM-based Data Augmentation on Downstream Tasks*. 2025. arXiv: 2502.04419 [cs.LG]. URL: <https://arxiv.org/abs/2502.04419>.
- [34] Yinheng Li et al. *Large Language Models in Finance: A Survey*. 2024. arXiv: 2311.10723 [q-fin.GN]. URL: <https://arxiv.org/abs/2311.10723>.
- [35] Minghao Liu et al. *Human and AI Perceptual Differences in Image Classification Errors*. 2025. arXiv: 2304.08733 [cs.CV]. URL: <https://arxiv.org/abs/2304.08733>.
- [36] Scott Lundberg and Su-In Lee. *A Unified Approach to Interpreting Model Predictions*. 2017. arXiv: 1705.07874 [cs.AI]. URL: <https://arxiv.org/abs/1705.07874>.
- [37] Fernando Martínez-Plumed et al. “CRISP-DM Twenty Years Later: From Data Mining Processes to Data Science Trajectories”. In: *IEEE Transactions on Knowledge and Data Engineering* 33.8 (2021), pp. 3048–3061. DOI: 10.1109/TKDE.2019.2962680.
- [38] Bertalan Mesko. “The ChatGPT (Generative Artificial Intelligence) Revolution Has Made Artificial Intelligence Approachable for Medical Professionals”. In: *Journal of medical Internet research* 25 (2023), e48392. DOI: <https://doi.org/10.2196/48392>.
- [39] Microsoft. *The LLM Latency Guidebook: Optimizing Response Times for GenAI Applications*. Accessed: 2025-08-01. 2025. URL: <https://techcommunity.microsoft.com/blog/azure-ai-services-blog/the-llm-latency-guidebook-optimizing-response-times-for-genai-applications/4131994>.
- [40] Hussein Mozannar et al. *Magnetic-UI: Towards Human-in-the-loop Agentic Systems*. 2025. arXiv: 2507.22358 [cs.AI]. URL: <https://arxiv.org/abs/2507.22358>.

- [41] Sonia Krishna Murthy, Tomer Ullman, and Jennifer Hu. “One fish, two fish, but not the whole sea: Alignment reduces language models conceptual diversity”. In: *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. Association for Computational Linguistics, 2025, pp. 11241–11258. DOI: 10.18653/v1/2025.naacl-long.561. URL: <http://dx.doi.org/10.18653/v1/2025.naacl-long.561>.
- [42] OpenAI. *ChatGPT: Optimizing language models for dialogue*. Accessed 2025-10-01. Nov. 2022. URL: <https://openai.com/index/chatgpt/>.
- [43] OpenAI. *Faulty Reward Functions*. Accessed: 2025-08-01. 2025. URL: <https://openai.com/index/faulty-reward-functions/>.
- [44] OpenAI. *Introducing OpenAI o3 and o4-mini*. Accessed: 2025-07-31. 2025. URL: <https://openai.com/index/introducing-o3-and-o4-mini/>.
- [45] OpenAI. *Introducing study mode*. Accessed: 2025-08-01. 2025. URL: <https://openai.com/index/chatgpt-study-mode/>.
- [46] OpenAI. *Learning to reason with LLMs*. Accessed: 2025-08-01. 2025. URL: <https://openai.com/index/learning-to-reason-with-l1ms/>.
- [47] OpenAI. *What were optimizing ChatGPT for*. Accessed: 2025-08-07. 2025. URL: <https://openai.com/index/how-we%27re-optimizing-chatgpt/>.
- [48] Max Peeperkorn et al. *Is Temperature the Creativity Parameter of Large Language Models?* 2024. arXiv: 2405.00492 [cs.CL]. URL: <https://arxiv.org/abs/2405.00492>.
- [49] AiMultiple Research. *AI Agents vs. Agentic AI Systems*. Accessed 2025-10-01. 2025. URL: <https://research.aimultiple.com/ai-agents-vs-agentic-ai/>.
- [50] Richard Rorty. *Contingency, irony, and solidarity*. Cambridge university press, 1989.
- [51] Stuart Russell and Peter Norvig. *Artificial Intelligence: A Modern Approach*. 4th. Pearson, 2021. ISBN: 978-0134610993.
- [52] Timo Schick et al. “Toolformer: Language Models Can Teach Themselves to Use Tools”. In: *ArXiv* abs/2302.04761 (2023). URL: <https://api.semanticscholar.org/CorpusID:256697342>.
- [53] Eike Schneiders et al. “Objection Overruled! Lay People can Distinguish Large Language Models from Lawyers, but still Favour Advice from an LLM”. In: *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. CHI 25. ACM, Apr. 2025, pp. 1–14. DOI: 10.1145/3706598.3713470. URL: <http://dx.doi.org/10.1145/3706598.3713470>.
- [54] Christoph Schröer, Felix Kruse, and Jorge Marx Gómez. “A Systematic Literature Review on Applying CRISP-DM Process Model”. In: *Procedia Computer Science* 181 (2021). CENTERIS 2020 - International Conference on ENTERprise Information Systems / ProjMAN 2020 - International Conference on Project MANagement / HCist 2020 - International Conference on Health and Social Care Information Systems and Technologies

- 2020, CENTERIS/ProjMAN/HCist 2020, pp. 526–534. ISSN: 1877-0509. DOI: <https://doi.org/10.1016/j.procs.2021.01.199>. URL: <https://www.sciencedirect.com/science/article/pii/S1877050921002416>.
- [55] Sara Stoudt, Váleri N. Vásquez, and Ciera C. Martinez. “Principles for data analysis workflows”. In: *PLOS Computational Biology* 17.3 (Mar. 2021). Ed. by Patricia M. Palagi, e1008770. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1008770. URL: <http://dx.doi.org/10.1371/journal.pcbi.1008770>.
 - [56] Yan Tao et al. “Cultural bias and cultural alignment of large language models”. In: *PNAS Nexus* 3.9 (Sept. 2024). Ed. by Michael Muthukrishna. ISSN: 2752-6542. DOI: 10.1093/pnasnexus/pgae346. URL: <http://dx.doi.org/10.1093/pnasnexus/pgae346>.
 - [57] Amos Tversky and Daniel Kahneman. “Judgment under Uncertainty: Heuristics and Biases: Biases in judgments reveal some heuristics of thinking under uncertainty.” In: *science* 185.4157 (1974), pp. 1124–1131.
 - [58] UNESCO. *Recommendation on the Ethics of Artificial Intelligence*. Accessed: 2025-08-01. 2021. URL: <https://www.unesco.org/en/legal-affairs/recommendation-ethics-artificial-intelligence>.
 - [59] Oxford University. *MSc in Social Data Science*. Accessed: 2025-08-01. 2025. URL: <https://www.ox.ac.uk/admissions/graduate/courses/msc-social-data-science>.
 - [60] Rodrick Wallace. “AI in the Real World”. In: *Carl von Clausewitz, the Fog-of-War, and the AI Revolution: The Real World Is Not A Game Of Go*. Cham: Springer International Publishing, 2018, pp. 1–45. ISBN: 978-3-319-74633-3. DOI: 10.1007/978-3-319-74633-3_1. URL: https://doi.org/10.1007/978-3-319-74633-3_1.
 - [61] University of Witwatersrand. *Wits ExplainableAI Lab*. Accessed: 2025-08-01. 2025. URL: <https://www.wits.ac.za/course-finder/postgraduate/science/msc-artificial-intelligence/>.
 - [62] Tongshuang Wu, Michael Terry, and Carrie Jun Cai. “AI Chains: Transparent and Controllable Human-AI Interaction by Chaining Large Language Model Prompts”. In: *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. CHI ’22. New Orleans, LA, USA: Association for Computing Machinery, 2022. ISBN: 9781450391573. DOI: 10.1145/3491102.3517582. URL: <https://doi.org/10.1145/3491102.3517582>.
 - [63] Weikai Xu et al. “LLM-Based Agents for Tool Learning: A Survey: W. Xu et al.” In: *Data Science and Engineering* (2025), pp. 1–31.
 - [64] Jingfeng Yang et al. *Harnessing the Power of LLMs in Practice: A Survey on ChatGPT and Beyond*. 2023. arXiv: 2304.13712 [cs.CL]. URL: <https://arxiv.org/abs/2304.13712>.
 - [65] Shunyu Yao et al. “React: Synergizing reasoning and acting in language models”. In: *International Conference on Learning Representations*. 2023.
 - [66] Yue Zhang et al. “Sirens Song in the AI Ocean: A Survey on Hallucination in Large Language Models”. In: *Computational Linguistics* (Sept. 2025), pp. 1–46. ISSN: 0891-2017. DOI: 10.1162/COLI.a.16. eprint: <https://doi.org/10.1162/COLI.a.16>.

- //direct.mit.edu/coli/article-pdf/doi/10.1162/COLI.a.16/2535477/coli.a.16.pdf. URL: <https://doi.org/10.1162/COLI.a.16.2535477>.
- [67] Zhengquan Zhang, Konstantinos Tsiakas, and Christina Schneegass. “Explaining the Wait: How Justifying Chatbot Response Delays Impact User Trust”. In: *Proceedings of the 6th ACM Conference on Conversational User Interfaces*. New York, NY, USA: Association for Computing Machinery, 2024. ISBN: 9798400705113. URL: <https://doi.org/10.1145/3640794.3665550>.
- [68] Kaitlyn Zhou, Dan Jurafsky, and Tatsunori Hashimoto. *Navigating the Grey Area: How Expressions of Uncertainty and Overconfidence Affect Language Models*. 2023. arXiv: 2302.13439 [cs.CL]. URL: <https://arxiv.org/abs/2302.13439>.

The Uses of Generative Artificial Intelligence for Cybersecurity in Organisations

Chipo Chidakwa¹ and Zainab Ruhwanya^{2[0000-0003-2339-7154]}

^{1,2} Cybersecurity and Privacy (CSPR) Research Group, Department of Information Systems,
University of Cape Town, Cape Town, South Africa

¹CHDCHI007@myuct.ac.za, ²zainab.ruhwanya@uct.ac.za
www.cspr.uct.ac.za

Abstract. Generative Artificial Intelligence (Gen AI) is an emerging technology that has the potential to influence cybersecurity within organisations. The transformative capabilities of Gen AI enable it to adapt to the rapidly evolving cyberspace in which organisations operate. This study represents primary efforts to understand Gen AI's capabilities for cybersecurity and its uses within organisations. The study was conducted through a literature review to investigate the discourse on the use of Gen AI for cybersecurity, based on published academic papers. The findings point to a discourse focused on the capabilities of Gen AI for cybersecurity within organisations, such as threat detection and automated testing, among others. The study also highlights the benefits associated with the usage of Gen AI for cybersecurity, as well as the concerns that arise within organisations. Further research will be necessary to determine employee perceptions of the use of Gen AI for cybersecurity within organisations.

Keywords: Generative Artificial Intelligence, Artificial Intelligence, Cybersecurity, Organisations.

1 Introduction

Organisations are increasingly integrating Generative Artificial Intelligence (Gen AI) into their cybersecurity workflows as the threat landscape and defensive tools evolve in tandem with the rapid development of Artificial Intelligence (AI). Deshpande & Gupta [1] note that there is a dynamic link between Gen AI and cybersecurity due to the protective and flexible capabilities of Gen AI that allow it to evolve and adjust according to the changing landscape of cyber threats. While it is widely acknowledged that eliminating all cyber threats is not feasible, Gen AI has exhibited capabilities that can augment detection, response, and user awareness, effectively minimising their subsequent risks [2, 3]. GenAI possesses the capability to efficiently process heterogeneous data and generate multimodal outputs [4]. When utilised in security contexts, these functionalities facilitate the translation of complex threat intelligence into stakeholder-specific summaries, enable realistic phishing and social engineering simulations for training purposes, and assist in identifying, explaining, or flagging manipulated media

and other malicious content across multiple languages and channels. Consequently, these applications serve to enhance an organisation's security posture.

However, Gen AI has exposed a new realm of risks within the cybersecurity landscape [5]. These risks include the creation of highly convincing phishing at scale, deep-fakes, automated social engineering, tool-assisted intrusion, and model-driven errors, such as hallucinations. [5–7]. The cyber risks that arise from the use of Gen AI for cybersecurity have an impact on an organisation's operations. Subsequently, efforts have been made to address these cyber risks, with some organisations implementing mitigation strategies [7]; however, the practice remains uneven.

Existing research has outlined the increased move towards incorporating Gen AI within organisational cybersecurity practices to leverage the opportunities that it presents. The implications of Gen AI usage for cybersecurity within organisations have exposed a gap in the understanding of the current uses of Gen AI for cybersecurity. This research aims to build on the existing research in this field and investigate the current uses of Gen AI for cybersecurity within an organisational context.

2 Background

The increase of AI advancements in the business world has led to the adoption of Gen AI within organisations for cybersecurity. Within organisations, cybersecurity is centred on safeguarding the hardware and software within cyberspace. [8]. Cybersecurity is seen to encompass the provision of capable resources and infrastructure to protect cyberspace by defending against and preventing cyber-attacks. AI enables machines to perform social and cognitive tasks, facilitating communication with other entities and processing high-level information [9]. Due to the advanced capabilities of AI, many organisations have begun incorporating emerging AI tools, such as Gen AI, into their organisational practices. Gen AI is an emerging technology capable of generating new, original content based on information provided to it and made available in various databases [10]. For example, the release of Gen AI tools such as ChatGPT and DALL-E by OpenAI has sparked growing interest in Gen AI within the AI community [11]. Interest has been drawn to the potential uses of Gen AI within organisations for cybersecurity and the subsequent implications.

The rapid rise of Gen AI has presented potential opportunities and threats in the cybersecurity landscape for organisations [5]. Existing research has highlighted a limited focus on research surrounding the uses and capabilities of Gen AI for cybersecurity within organisations, to leverage the potential opportunities and mitigate the possible threats that it presents. This gap implies the need to understand the uses of Gen AI for cybersecurity within organisations. The following research question has been derived to address this research topic and the main objective of this research:

Research Question: What are the uses of Gen AI for cybersecurity in organisations? With the primary objective of this research being to explore the applications of Gen AI for cybersecurity in organisations, this study assessed papers found through a literature search that investigate the intersection of Gen AI and cybersecurity within organisations. The capabilities of Gen AI for cybersecurity, its benefits, and drawbacks for

usage within organisations were investigated. The section below outlines the research methodology used to arrive at the findings.

3 Research Design

This research was conducted as a systematic literature review (SLR). The research protocol was guided by Okoli [12], with the process beginning by deriving keywords to be used for searching relevant literature in academic databases and other online platforms. Inclusion and exclusion criteria were established and used to screen the identified papers and exclude irrelevant ones. The remaining papers were then analysed, and key themes were identified and categorised for the findings.

3.1 Literature Search and Selection Strategy

As the focus of the study is Gen AI usage for cybersecurity in organisations, the key search terms used to source the relevant publications were: (“Generative Artificial Intelligence” OR “Gen AI”) AND (“CYBERSECURITY”) AND (“ORGANISATIONS.”) The selected search terms were inputted into the Scopus and Web of Science databases as the two databases typically cover a wide range of publications relevant to any information systems or technology related discipline [12]. Additional publications were identified through a comprehensive literature search and snowball sampling. This “hybrid” technique was utilised for this research as the area of focus is new and continuously being updated, with added information being uncovered and made available at a rapid rate. Wohlin et al. [13] highlight that this technique is common in SLRs to improve the quantity and enhance the quality of the publications that are analysed. The publications found were then scrutinised against the inclusion and exclusion criteria outlined in Table 1.

Table 1. Inclusion and Exclusion Criteria

Criteria	Include	Exclude
Year	Published between 2021 and 2025. Due to the rapidly evolving nature of Gen AI, it is important for recent sources to be screened and analysed	Research published before 2021
Paper Focus	Exploring the uses of Gen AI for Cybersecurity in organisations	Focus is solely on the technical elements of Cybersecurity
Language	Written in English	Not written in English
Area of Focus	<ul style="list-style-type: none"> • Capabilities of Gen AI • Benefits of the use of Gen AI in an organisational context • Concerns surrounding the uses of Gen AI in an organisational context • Cybersecurity practices within an organisation that are impacted by AI 	<ul style="list-style-type: none"> • No focus on the capabilities of Gen AI in the context of organisational cybersecurity • Focus on the technical elements of Gen AI and the different Gen AI models

3.2 Screening Process

This study followed the screening process outlined by Okoli [12] as a guide for selecting the appropriate papers. The study employed a two-phase screening approach, which began with searching for literature relevant to the topic using the defined search terms for this study, and then assessed the titles, abstracts, and keywords from the identified studies. Following this, the next phase screened the papers that had been narrowed down in Phase 1 and examined the entire text of the literature to identify studies that contained relevant information to the study focus. By following this screening process, the study gathered a relevant dataset for analysis.

A Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) flowchart diagram is used to outline the different phases of the systematic review of the literature [14]. Figure 1 outlines the PRISMA flowchart diagram, illustrating the data collection and screening process undertaken for this literature search.

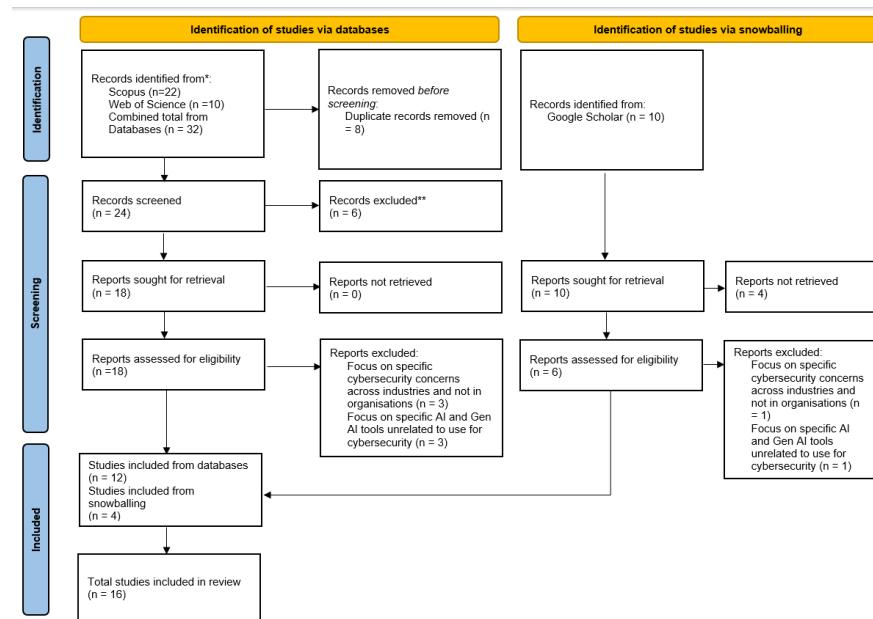


Fig. 1. PRISMA Diagram

The combined sourced papers were then analysed to derive the relevant insights for this research. These papers are listed in table 1 below.

Table 2. Systematic Literature Review Papers

	Paper Title	Paper Title
[2]	Cyber Attack Prediction: From Traditional Machine Learning to Generative Artificial Intelligence	[15] Role of Artificial Intelligence based Chat Generative Pre-trained

[16]	Lateral Phishing With Large Language Models: A Large Organization Comparative Study	[3]	Transformer (ChatGPT) in Cyber Security
[17]	The Impact of Generative AI and LLMs on the Cybersecurity Profession.	[18]	Examine the Role of Generative AI in Enhancing Threat Intelligence and Cyber Security Measures
[1]	GenAI in the Cyber Kill Chain: A Comprehensive Review of Risks, Threat Operative Strategies and Adaptive Defense Approaches.	[19]	Generative AI for Cyber Security: Analyzing the Potential of ChatGPT, DALL-E, and Other Models for Enhancing the Security Space
[20]	Security and Privacy Perspectives on Using ChatGPT at the Workplace: An Interview Study	[21]	Bridging knowledge gap: the contribution of employees' awareness of AI cyber risks comprehensive program to reducing emerging AI digital threats.
[22]	Enhancing Cyber Security Enhancement Through Generative AI	[23]	AI for cyber-security risk: harnessing AI for automatic generation of company-specific cybersecurity risk profiles
[24]	Generative AI in Cybersecurity	[25]	What The Phish! Effects of AI on Phishing Attacks and Defense
[26]	Cyber Security Issues and Challenges Related to Generative AI and ChatGPT	[27]	The paradigm of hallucinations in AI-driven cybersecurity systems: Understanding taxonomy, classification outcomes, and mitigations
			GAI-Driven Offensive Cybersecurity: Transforming Pentesting for Proactive Defence

3.3 Literature Analysis Procedure

To analyse the extracted data, a thematic analysis following the six-step framework outlined by Braun and Clarke [28] was employed to identify, examine, and document the uncovered themes and concepts. As outlined by Maguire and Delahunt [29], a thematic analysis is valuable for identifying and highlighting the key themes and their interrelationships in a meaningful way. The 16 papers that were included for analysis were input into NVIVO software for coding to conduct a comprehensive analysis of the data gathered [30, 31]. An initial set of 208 codes was generated from an overall analysis of the papers. From this, themes were developed based on the recurring patterns identified, and the relevance and accuracy of the defined themes were then assessed, resulting in 3 main themes defined from the papers for interpretation. The following themes were outlined for this study: Understanding Gen AI and cybersecurity, the benefits of Gen AI usage for cybersecurity, and the drawbacks related to Gen AI usage for cybersecurity.

4 Results and Discussions

The findings from this study provide an overview of the themes which reflect the intersection between Gen AI and cybersecurity within organisations. This study explored the capabilities and benefits of Gen AI for cybersecurity, as well as the associated concerns and challenges.

4.1 Understanding Gen AI for Cybersecurity

With the rapid rise of AI and its growing usage within organisations, it is important to understand the emerging technologies that fall under it and their growing usage within organisations. AI is defined by Capodieci et al. [17]p.448 as “the field of study that analyses the creation of intelligent machines”. While Prasad et al. [15]p.107 recognise AI as a “multidisciplinary technology” that can emulate human behaviours and tasks more efficiently, store data, make decisions, and promote machine learning. The advanced nature of AI as a technology has seen it being deployed in many contexts, from organisations to individuals’ lives. AI offers several advantages for addressing security concerns, including detecting network vulnerabilities, analysing large datasets to identify suspicious activities, and automating elements of incident response to enable timely countermeasures [3]. The field has evolved over the years, transforming technological capabilities. Schreiber & Schreiber [19] highlight the shift from rule-based systems to the more recent subset, GenAI, which can produce original content and be utilised for security. AI technologies continue to advance, with GenAI and related tools developing rapidly [26].

GenAI models exhibit transformative capabilities that can impact an organisation’s cybersecurity space. Large language models (LLMs) fall under GenAI and use deep learning and statistical methods to analyse vast amounts of data and predict patterns [2, 17]. These models can generate realistic content and assist with tasks that improve aspects of an organisation’s cybersecurity profile [3, 16]. Their value has been observed in organisational settings, particularly for cybersecurity. In some cases, GenAI models are developed, fine-tuned, or privately deployed to maintain control over data handling and governance, while others rely on enterprise-grade, vendor-hosted options with contractual data protections [17, 20].

However, GenAI has also made it more difficult for organisations to detect malicious threats as attack techniques evolve. Adversaries can leverage GenAI to craft persuasive phishing and social-engineering campaigns, generate deepfakes, and automate parts of exploitation, which can outpace some legacy controls [23, 24]. Such capabilities risk eroding trust, misleading employees, and prompting actions that compromise privacy settings and safeguards [2, 20, 21]. The evolving nature of cyber threats underscores the need for enhanced, AI-aware defences that specifically address GenAI-enabled risks [19]. As such, GenAI presents a double-edged sword for organisational cybersecurity; its defensive benefits are substantial, but so too are the new attack surfaces and modes of deception it enables.

4.2 Benefits of Gen AI Usage for Cybersecurity

Gen AI can analyse data, identify threat patterns, and surface insights that help assess the effectiveness of an organisation's security posture. By using predictive capabilities on large datasets, Gen AI models can improve threat detection and inform security decisions [3, 24]. Organisations also utilise these models to generate realistic scenarios and synthetic data to simulate testing environments and evaluate the effectiveness of their threat detection and risk analysis approaches [3, 18]. Furthermore, adversarial testing is used to probe the vulnerabilities of Gen AI-enabled systems and to develop more robust models and controls [18]. AI-enabled security tooling can augment monitoring and adapt scoring or prioritisation as risks evolve in organisations. By using Gen AI to develop and enhance cybersecurity measures, organisations can strengthen protections while leveraging the opportunities these technologies present. Overall, AI-generated security systems can combine the benefits of Gen AI for cybersecurity by effectively identifying and adapting to evolving cyber risks in organisations [1, 21].

Threat intelligence and detection have improved in efficiency with the introduction of Gen AI for cybersecurity [21]. Gen AI models can protect against cyber-attacks by identifying potential threats and providing updated security insights that support risk mitigation [3, 22]. These systems can also automate elements of incident response, such as enrichment, triage, and playbook steps, typically with human-in-the-loop oversight, to contain breaches more quickly [1]. This increases the chances of minimising the impact of attacks and containing risk efficiently. Where appropriate, automation enables timely adjustments to countermeasures; however, full real-time autonomy is uncommon and is usually governed by policy [2].

Cybersecurity experts can leverage Gen AI models to analyse existing datasets and identify potential threats [3]. The same capabilities can support system response design and employee awareness training. For example, Gen AI can generate realistic training materials or simulated phishing content to increase vigilance and reduce exposure to sophisticated attacks, provided outputs are reviewed for accuracy and context [3, 16, 19]. Cybersecurity testing profiles are generated by these models to familiarize employees with common trends in cyber-attacks. Therefore, the usage of Gen AI to ensure that employees are aware of the evolving cyber threat landscape and to train their system responses enhances the cybersecurity of organisations.

Gen AI has been shown to provide beneficial enhancements to organisational cybersecurity across business domains[21]. Many organisations are implementing Gen AI-enabled tools within their security programmes. Used appropriately, Gen AI can enhance defensive measures by helping to detect and mitigate cyberattacks[2, 3, 22]. In summary, key benefits include improved threat detection, partial automation of incident response, analysis of large datasets to create testing scenarios, and timely adjustments to countermeasures.

4.3 Drawbacks of Gen AI Usage for Cybersecurity

Despite the transformative capabilities of Gen AI for cybersecurity within organisations, attackers are also leveraging these tools to lower barriers to entry and to scale and

customise cyber-attacks [24]. This dynamic contributes to greater sophistication in attack campaigns, even if overall incident rates are influenced by many factors. Organisations face a range of AI-related risks, including large-scale phishing and social engineering, the spread of misinformation, deepfakes, AI-assisted malware development and polymorphism, automated reconnaissance and exploitation, data manipulation and poisoning, prompt injection, and adversarial prompt attacks [17, 19, 21–23, 25, 26]. In parallel, Gen AI deployments introduce system-level risks such as hallucinations that can mislead analysts or users, inadvertent leakage of sensitive data through prompts or outputs, and integrity threats to models and data pipelines (e.g., poisoning or inversion) [17, 22].

Ethical Concerns and Limited Employee Awareness

Gen AI models are trained on large datasets and require this data to generate functional outputs. In practice, employees may inadvertently input sensitive information into Gen AI tools when seeking help to draft security strategies or analyse configurations, which can expose privacy risks if data-handling settings are not well controlled [26]. Many employees have limited experience with AI applications and are often unaware of these risks [19, 20]. As such, confidential information may be unintentionally shared via Gen AI tools. Where retention or model training on user inputs is permitted, sensitive details could be revealed through outputs or logs, increasing the risk of disclosure to unauthorised recipients [17, 22]. Furthermore, some vendor-hosted Gen AI services may retain prompts and outputs by default (unless enterprise controls disable this), which, together with misconfigurations or third-party integrations, can create leakage pathways that compromise the privacy and integrity of organisational data.

Ethical concerns also arise around data accuracy and the provenance of information used by Gen AI systems. Models can inherit biases from their training data, leading to unintentional unfair outcomes [17, 26]. Limited interpretability makes such biases harder to detect and explain, which can undermine transparency and decision quality in cybersecurity use cases [25]. While Gen AI providers implement policies and safety filters, these safeguards can be bypassed through “jailbreaking” prompts, leaving organisations exposed to misuse if internal controls are weak [24]. Because Gen AI content is derived from heterogeneous datasets, comprehensive accuracy checks are challenging. Many organisations further lack clear ethical guidelines for acceptable use, data classification, and review procedures, increasing exposure to data risks [20, 27]. To mitigate these issues, organisations should aim for Gen AI-generated outputs that are reviewed for bias and transparency and governed by explicit ethical policies, covering retention settings, no-training modes, access controls, and human-in-the-loop review to improve the protection and security of sensitive data [21].

Malicious Actors and Fraudsters

Fraudsters and threat actors are leveraging Gen AI to avoid detection and to launch sophisticated, targeted cyber-attacks more efficiently [1]. Malicious actors have identified vulnerabilities in AI-enabled systems that can be exploited through automation of reconnaissance and exploitation steps and through adversarial techniques (e.g., prompt

injection and evasion) [2, 22]. In this way, Gen AI can be misused to generate convincing lures, tailor payloads, and scale attacks, creating defensive challenges for organisations that adopt these tools [21, 24]. The overall impact varies across sectors and environments, but there is clear evidence of increasing sophistication and scale in some attack campaigns that make use of Gen AI [17]. As such, companies that employ Gen AI for cybersecurity are exposed to malicious actors who may use the same technologies to seek unauthorised access to confidential information and to execute cyber-attacks, underscoring the need for strong governance, monitoring, and guardrails.

Deepfakes and Phishing Attacks

Deepfake technology enables malicious actors to generate convincing synthetic content to deceive individuals and spread [17, 19, 26]. Using deepfakes, attackers can misrepresent themselves as trusted employees or executives to induce fraudulent actions under the appearance of legitimate requests [2, 16]. Employees may become susceptible to such attacks because of the increasing realism of deepfakes, which makes it difficult to distinguish authentic from fabricated content and can accelerate the spread of misinformation [19, 24]. Deepfakes can therefore misdirect staff through identity impersonation and fraud, exposing organisations to privacy and integrity risks and compromising security processes [18]. The impact of deepfake technology on organisational cybersecurity highlights the potential limitations of Gen AI for cybersecurity when governance and verification controls are weak.

Phishing attacks can be socially engineered by cyber attackers using Gen AI models to deceptively obtain access to private and sensitive data from employees [2, 17, 26]. Attackers can enhance their tactics with AI to improve targeting, tone, and timing, posing material risks to an organisation's cybersecurity posture [22]. Such misuse allows adversaries to craft sophisticated messages that appear legitimate and manipulate recipients into disclosing sensitive information or executing risky actions [24]. Detecting AI-generated phishing emails is challenging for employees, especially without proper training and robust layered controls, because the messages are often persuasive and consistent with internal styles [23].

4.4 Conceptual Framework

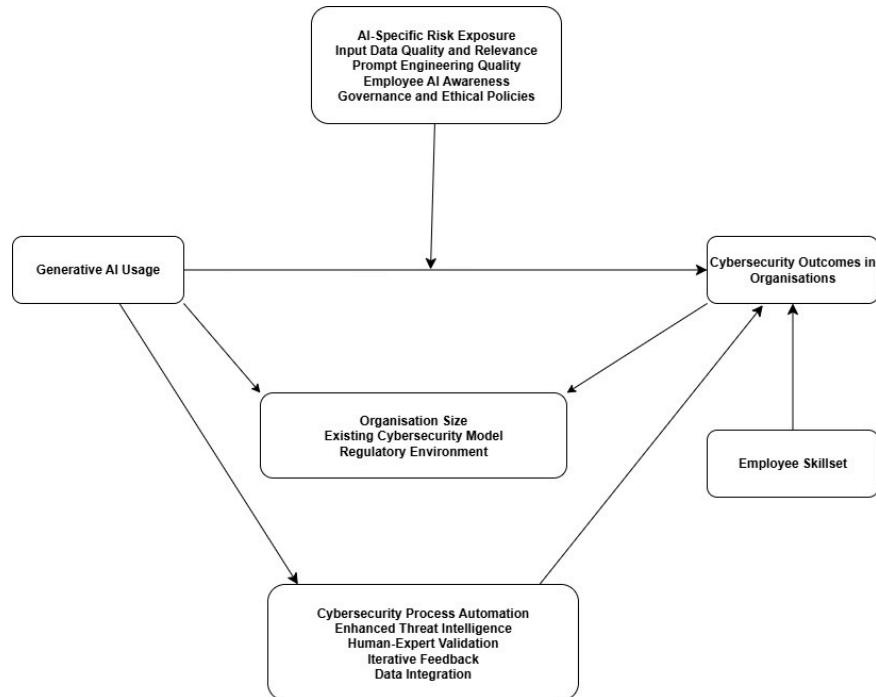


Fig. 2. Conceptual framework linking GenAI usage to organisational cybersecurity outcomes

The figure shows how Gen AI usage relates to cybersecurity outcomes in organisations. Outcomes refer to the ability to detect, respond to, and prevent threats. Gen AI can influence outcomes directly. It can also do so by improving security processes such as process automation, enhanced threat intelligence, human-expert validation, iterative feedback, and data integration. The strength of this influence depends on several factors: AI-specific risk exposure, input data quality and relevance, prompt-engineering quality, employee AI awareness, and governance and ethical policies. Contextual factors can shape both Gen AI adoption and outcomes. These include organisation size, the current cybersecurity model, and the regulatory environment. Employee skillset also affects outcomes and is shown separately in the diagram

5 Conclusion and Future Research Directions

This study provided an understanding of Gen AI and its capabilities for cybersecurity within organisations. Gen AI can analyse large datasets, identify threat patterns, and surface insights that inform security decisions. Additionally, Gen AI models can adapt to the evolving threat landscape and automate incident responses, typically with human

oversight, to ensure an effective response to attacks and minimise the impact of cyberattacks. These models also provide valuable insights for enhancing cybersecurity systems and simulate cyberattacks to raise awareness of potential threats across the workforce.

The study also found important drawbacks. Attackers can use Gen AI to run deceptive campaigns, for example deepfakes and highly tailored phishing, that exploit employee vulnerabilities. A limited understanding of Gen AI among staff can lead to the accidental sharing of sensitive information with large language models, thereby weakening data security and privacy. Because Gen AI learns from varied datasets, it may reproduce biases and create integrity risks if outputs are not reviewed. Gaps in ethical guidance and data-handling rules further reduce organisational control over what is entered into tools and over the accuracy of the outputs. Organisations need to recognise these concerns and put mitigation strategies in place so that benefits can be realised responsibly.

Given this double-edged nature, many organisations struggle to keep pace with rapid Gen AI advances, which can leave them exposed to evolving threats. The practical task is to strengthen governance, configure tools correctly, maintain human review, and build targeted awareness so that Gen AI supports defence rather than creating avoidable risk.

This paper outlined the benefits and drawbacks of Gen AI for organisational cybersecurity based on published academic literature. While prior work covers many practical uses, there is limited coverage of how employees perceive Gen AI in security work, even though those perceptions can create opportunities or additional risks. Future research should examine these perceptions to provide a more complete understanding of Gen AI use for cybersecurity in organisations. Two research questions follow: How do perceptions and uses of Gen AI differ between SMEs and large enterprises, and what does that mean for outcomes? Which security awareness and review approaches help staff detect AI-enabled phishing and deepfakes in practice?

References

1. Deshpande AS, Gupta S (2023) GenAI in the Cyber Kill Chain: A Comprehensive Review of Risks, Threat Operative Strategies and Adaptive Defense Approaches. In: 2023 IEEE International Conference on ICT in Business Industry & Government (ICTBIG). IEEE, Indore, India, pp 1–5
2. Ankalaki S, Atmakuri AR, Pallavi M, Hukkeri GS, Jan T, Naik GR (2025) Cyber Attack Prediction: From Traditional Machine Learning to Generative Artificial Intelligence. IEEE Access 13:44662–44706. <https://doi.org/10.1109/ACCESS.2025.3547433>
3. Saddi VR, Gopal SK, Mohammed AS, Dhanasekaran S, Naruka MS (2024) Examine the Role of Generative AI in Enhancing Threat Intelligence and Cyber Security Measures. In: 2024 2nd International Conference on Disruptive Technologies (ICDT). IEEE, Greater Noida, India, pp 537–542
4. Prasad Agrawal K (2024) Towards Adoption of Generative AI in Organizational Settings. Journal of Computer Information Systems 64:636–651. <https://doi.org/10.1080/08874417.2023.2240744>
5. Neupane S, Fernandez IA, Mittal S, Rahimi S (2023) Impacts and Risk of Generative AI Technology on Cyber Defense

6. Gupta M, Akiri C, Aryal K, Parker E, Praharaj L (2023) From ChatGPT to ThreatGPT: Impact of Generative AI in Cybersecurity and Privacy. *IEEE Access* 11:80218–80245. <https://doi.org/10.1109/ACCESS.2023.3300381>
7. Singla A, Sukharevsky A, Yee L, Chui M, Hall B (2025) The State of how Organizations are Rewiring to Capture Value
8. Jada I, Mayayise TO (2024) The impact of artificial intelligence on organisational cyber security: An outcome of a systematic literature review. *Data and Information Management* 8:100063. <https://doi.org/10.1016/j.dim.2023.100063>
9. Abbass H (2021) Editorial: What is Artificial Intelligence? *IEEE Trans Artif Intell* 2:94–95. <https://doi.org/10.1109/TAI.2021.3096243>
10. Teo ZL, Quek CWN, Wong JLY, Ting DSW (2024) Cybersecurity in the generative artificial intelligence era. *Asia-Pacific Journal of Ophthalmology* 13:100091. <https://doi.org/10.1016/j.apjo.2024.100091>
11. Byrne MD (2023) Generative Artificial Intelligence and ChatGPT. *Journal of PeriAnesthesia Nursing* 38:519–522. <https://doi.org/10.1016/j.jopan.2023.04.001>
12. Okoli C (2015) A Guide to Conducting a Standalone Systematic Literature Review. *CAIS* 37:. <https://doi.org/10.17705/1CAIS.03743>
13. Wohlin C, Kalinowski M, Romero Felizardo K, Mendes E (2022) Successful combination of database search and snowballing for identification of primary studies in systematic literature studies. *Information and Software Technology* 147:106908. <https://doi.org/10.1016/j.infsof.2022.106908>
14. PRISMA 2020 flow diagram. In: PRISMA statement. <https://www.prisma-statement.org/prisma-2020-flow-diagram>. Accessed 30 Sept 2025
15. Prasad SG, Sharmila VC, Badrinarayanan MK (2023) Role of Artificial Intelligence based Chat Generative Pre-trained Transformer (ChatGPT) in Cyber Security. In: 2023 2nd International Conference on Applied Artificial Intelligence and Computing (ICAAIC). IEEE, Salem, India, pp 107–114
16. Bethany M, Galiopoulos A, Bethany E, Bahrami Karkevandi M, Beebe N, Vishwamitra N, Najafirad P (2025) Lateral Phishing With Large Language Models: A Large Organization Comparative Study. *IEEE Access* 13:60684–60701. <https://doi.org/10.1109/ACCESS.2025.3555500>
17. Capodieci N, Sanchez-Adames C, Harris J, Tatar U (2024) The Impact of Generative AI and LLMs on the Cybersecurity Profession. In: 2024 Systems and Information Engineering Design Symposium (SIEDS). IEEE, Charlottesville, VA, USA, pp 448–453
18. Sai S, Yashvardhan U, Chamola V, Sikdar B (2024) Generative AI for Cyber Security: Analyzing the Potential of ChatGPT, DALL-E, and Other Models for Enhancing the Security Space. *IEEE Access* 12:53497–53516. <https://doi.org/10.1109/ACCESS.2024.3385107>
19. Schreiber A, Schreiber I (2024) Bridging knowledge gap: the contribution of employees' awareness of AI cyber risks comprehensive program to reducing emerging AI digital threats. *ICS* 32:613–635. <https://doi.org/10.1108/ICS-10-2023-0199>
20. Kimbel A, Glas M, Pernul G (2025) Security and Privacy Perspectives on Using ChatGPT at the Workplace: An Interview Study. In: Clarke N, Furnell S (eds) *Human Aspects of Information Security and Assurance*. Springer Nature Switzerland, Cham, pp 184–197
21. Schreiber A, Schreiber I (2025) AI for cyber-security risk: harnessing AI for automatic generation of company-specific cybersecurity risk profiles. *ICS*. <https://doi.org/10.1108/ICS-08-2024-0177>
22. Krishnamurthy O (2023) Enhancing Cyber Security Enhancement Through Generative AI. Vol No

23. Shreyas Kumar, Anisha Menezes, Sarthak Giri, Srujan Kotikela (2024) What The Phish! Effects of AI on Phishing Attacks and Defense. ICAIR 4:218–226. <https://doi.org/10.34190/icair.4.1.3224>
24. Metta S, Chang I, Parker J, Roman MP, Ehuan AF (2024) Generative AI in Cybersecurity
25. Sood AK, Zeadally S, Hong E (2025) The paradigm of hallucinations in AI-driven cybersecurity systems: Understanding taxonomy, classification outcomes, and mitigations. Computers and Electrical Engineering 124:110307. <https://doi.org/10.1016/j.compeleceng.2025.110307>
26. Pasupuleti R, Vadapalli R, Mader C (2023) Cyber Security Issues and Challenges Related to Generative AI and ChatGPT. In: 2023 Tenth International Conference on Social Networks Analysis, Management and Security (SNAMS). IEEE, Abu Dhabi, United Arab Emirates, pp 1–5
27. Zaydi M, Maleh Y (2025) GAI-Driven Offensive Cybersecurity: Transforming Pentesting for Proactive Defence: In: Proceedings of the 11th International Conference on Information Systems Security and Privacy. SCITEPRESS - Science and Technology Publications, Porto, Portugal, pp 426–433
28. Braun V, Clarke V (2006) Using thematic analysis in psychology. Qualitative Research in Psychology 3:77–101. <https://doi.org/10.1191/1478088706qp063oa>
29. Maguire M, Delahunt B Doing a Thematic Analysis: A Practical, Step-by-Step Guide for Learning and Teaching Scholars.
30. Bhattacherjee A (2012) Social Science Research: Principles, Methods, and Practices. Global Text Project, Place of publication not identified
31. Leech NL, Onwuegbuzie AJ (2011) Beyond constant comparison qualitative data analysis: Using NVivo. School Psychology Quarterly 26:70–84. <https://doi.org/10.1037/a0022711>

Part VI

**Volume II: Responsible and
Ethical AI**

Otheroids or Anthropomorphism? An Empathy-Based Approach to Artificial Agents

Abootaleb Safdari¹

¹ University of Bremen, Bibliothekstraße 1, 28359 Bremen, Germany
asafdar@uni-bremen.de

Abstract. The increasing presence of artificial agents (AAs) in everyday life has foregrounded a profound paradox in human-machine interaction: while people instinctively engage with AAs as if they possess consciousness and emotions, they often intellectually deny these very attributions. This contradiction has led to a dominant research paradigm, which this paper terms the "deception strategy," that dismisses such empathic behavior as a fallacy rooted in anthropomorphic illusions. This paper argues that this view is flawed, as it relies on (1) a rigid ontological divide between humans and machines and a (2) simplistic distinction between appearance and reality. Drawing from phenomenologically inspired enactivism, this paper proposes an alternative framework that reinterprets these empathic responses not as deceptive projections, but as the constitutive elements of a new form of social relation. By introducing the concept of the "otheroid," this paper offers a novel category for artificial entities that are experienced as neither fully human nor purely mechanical, thereby embracing the dynamic, reciprocal, and embodied nature of our interactions in an increasingly technologically mediated world.

Keywords: Anthropomorphism, Deception, Empathy, Phenomenology, Enactivism; Otheroid

1 Introduction

Our new brave world is increasingly populated by artificial agents (AAs), from robotic companions and digital assistants to autonomous systems in healthcare and industry. As these AAs become more sophisticated, human interaction with them has evolved beyond simple utility to encompass social and emotional dimensions. Studies have shown that people form strong, human-like bonds with these agents. For instance, research on the robotic seal PARO, used for dementia care, found that patients treated it as a social companion, a source of happiness, and even a conversation partner (Hung et al., 2021). Similarly, users of the robot Vector during quarantine described it as a family member and companion, engaging in shared activities like dancing (Odekerken-Schröder et al., 2020). This deeply ingrained human-like engagement with AAs has become a central topic of study across various disciplines, yet it suffers from significant conceptual ambiguity (Damholdt et al., 2023; Thellman et al., 2022).

This paper adopts the term anthropomorphism to refer to this phenomenon of attributing human-like qualities and mental states to AAs (Breazeal, 2002). The literature reveals a core paradox at the heart of this concept: while people readily anthropomorphize AAs during real-time, embodied encounters, they just as readily deny any true animacy or subjectivity to them upon detached, cognitive reflection. This apparent contradiction has led many researchers to adopt what this paper identifies as a deception strategy, arguing that such anthropomorphic behaviors are fundamentally illusory and should be discouraged. This paper will argue that this denial strategy is a conceptual dead end, rooted in a problematic adherence to a strict human-machine dualism and a simplistic distinction between appearance and reality. It will then propose a positive alternative: an empathy-based framework grounded in phenomenological enactivism and relationalism that reframes empathic human-AA interactions as genuine and constitutive of a new kind of social relation, which we describe with the new concept of the otheroid.

2 The Paradox and Deception Strategy

The paradox of human-AA interaction lies in the disconnect between implicit and explicit mind ascription. As Epley et al. (2007) noted, individuals with less time or cognitive resources for "effortful correction" are more likely to exhibit anthropomorphic behavior. This has been confirmed by empirical studies. In a study by Banks (2020), participants implicitly mentalized robots in a manner similar to humans during Theory of Mind (ToM) tests, yet explicitly denied that the robots had minds when directly asked. This suggests that the brain automatically processes social cues from AAs, but a reflective, conscious process overrides this intuitive response with the rational understanding that the robot is "just a machine." A similar finding by Fussell et al. (2008) showed that while participants used human-like language and attributed social traits to a robotic interviewer in spontaneous, real-time descriptions, they later explicitly denied that the robot possessed emotions or moods in a post-task survey. The data indicates that people engage in automatic social cognition with robots, but this is a far cry from a reflective belief that these agents have minds.

This paradoxical dynamic has led many scholars to conclude that attributing subjectivity to AAs is fundamentally mistaken—a form of deception. The deception strategy, as articulated by Matthias (2015), Placani (2024), Sharkey & Sharkey (2012, 2021), Sparrow & Sparrow (2006) and Winkle et al. (2021), posits that anthropomorphic design, emotional cues, and natural language in robots are deceptive techniques. They argue that these features exploit human social responsiveness without any corresponding internal states in the robot. This is seen as a form of "passive deception" arising from a performative gap between a robot's appearance and its actual function (2015). This view implies a duty to "see the world as it is" and avoid the "sentimentality" of believing an electronic toy is a friend (Sparrow & Sparrow, 2006, p. 155).

This widespread rejection of anthropomorphism as deception is, however, built upon two problematic assumptions. The first is a rigid ontological divide between humans (sentient, subjective beings) and AAs (passive, inanimate objects). The second is a

metaphysical separation between reality and appearance, where the robot's "true nature" as a machine is hidden beneath a deceptive, human-like facade. Critically, both assumptions fall apart under closer scrutiny of human-AA interactions.

3 A Critique of the Deception Strategy

As it has been mentioned, the "deception strategy" for understanding anthropomorphism is based on two faulty ideas. First, A strict divide between humans and machines that sees humans as having true subjectivity and machines as passive, lifeless objects. Thus attributing human traits to a machine is considered a mistake. Second, A separation of appearance and reality that suggests that a robot's human-like appearance is a fake illusion that hides its true mechanical nature. I believe both of these assumptions are problematic. Before addressing them in the next section, I would like to raise a preliminary concern about this strategy. First, even knowledgeable experts like Sherry Turkle (2011, p. 84) who know that a robot is "just code," still find themselves reacting to it as if it were a person. This proves that anthropomorphic behavior is not simply a mistake that can be corrected with knowledge. Second, The Roomba vacuum, a clearly mechanical and non-humanoid device, still inspires "Roombarization" in its users, who name it and express concern for it. This shows that anthropomorphism isn't just a response to a deceptive, human-like appearance; it's a deeper, more fundamental aspect of human-machine interaction.

3.1 Human-Machine Divide

In *Facing Gaia* (2015), Bruno Latour critiques entrenched dichotomies such as nature/culture, human/nonhuman, and animate/inanimate. He shows that natural entities are often treated as agents in practice—even by those who deny them agency. For example, the Mississippi River is described as "choosing" its path or "defying" dams. Such language reflects a practical recognition that the river exerts force, resists control, and requires negotiation, much like a political actor. Latour urges abandoning these binaries in favor of a more open notion of the "world" or "pluriverse" (after William James), which embraces the diversity and irreducibility of beings and resists fixed categories.

In this view, we live in a *metamorphic zone*—a space of constant transformation where boundaries between entities are fluid. Here, humans, animals, machines, and natural systems continually shape one another. It is within the metamorphic zone that anthropomorphism reveals itself not as a one-way projection from humans onto machines, but as a reciprocal dynamic, in which machines also shape and reconfigure human perception and embodiment. In *Seeing Like a Rover*, Janet Vertesi introduces the concept of technomorphism to describe this reversal of perspective, where humans begin to adapt to and internalize the robotic body's way of engaging with the world (2015). Instead of simply imagining the robot as human-like, scientists and engineers working

on the Mars Rover missions learn to "see like a rover"—they embody its sensory constraints, operational logic, and limited range of motion.

Apart from these broader observations, we can narrow our focus to a specific subzone within the larger metamorphic zone—namely, the human–AA subzone—and develop it in a more systematic and conceptually rigorous way. To this aim, relationalism offers a more systematic way to understand human-AA interactions. This philosophical view argues that relationships between entities are fundamental, meaning an entity's nature is defined by its interactions, not by its fixed, intrinsic properties. This approach challenges traditional views that see subjects and objects as separate. In the context of robotics and AI, relationalism suggests that these systems are best understood and evaluated through their relationships with humans, society, and the environment, rather than as isolated tools (Coeckelbergh, 2012b; Darling, 2021; Gellers, 2021; Gunkel, 2023; Jones, 2013; Puzio, 2024).

3.2 Relationalism

Coeckelbergh (2010, 2012b, 2022) argues that traditional ontological, properties-based approaches treat moral status as something grounded in an entity's intrinsic characteristics. On this view, entities qualify for moral consideration if they possess certain defining traits: rationality, consciousness, sentience, or the ability to experience pleasure and pain. He then argues that this framework is fundamentally flawed in three ways:

1. **Epistemological problem** – There is no agreement on what Q should be; philosophers and ethicists disagree on which property is decisive.
2. **Detection problem** – Even if we agreed on Q , we often cannot reliably detect it, especially when it involves internal states.
3. **Continuum problem** – Most candidate properties come in degrees, making moral boundaries arbitrary and exclusionary.

In place of this intrinsic-property model, Coeckelbergh advances a *relational* approach. Here, moral status does not reside *in* the entity as an inherent essence. Instead, it emerges through lived interaction—how the entity is experienced, engaged with, and situated within broader social, cultural, linguistic, and spatial contexts. This “phenomenological–transcendental” shift reframes moral status as a product of relational dynamics rather than static ontological facts.

David Gunkel (2012, 2023), like Coeckelbergh, critiques the traditional approach to robot ethics, which frames the debate around dualist dichotomies like "person/thing." He argues this debate is a dead end because both sides rely on a flawed logical framework: that an entity must possess a certain quality (Q) to be considered a person. Similarly, Gunkel identifies three problems with this logic: there's no agreement on what "Q" is (the determination problem), no way to reliably detect it (the epistemological problem), and any judgment about its presence becomes a subjective decision (the decision problem). To escape this predicament, Gunkel proposes a radical shift: ethics precedes ontology. Instead of a robot's inherent nature dictating how we should treat it, he argues that our ethical decision to respond to it—drawing on Levinasian ethics—is what constitutes its status. In this view, how we treat a robot determines what it is, rather than the other way around.

3.3 Appearance and Reality

Mark Coeckelbergh challenges the "deception" narrative of anthropomorphism by arguing it's based on a flawed Platonic metaphysics that creates a strict divide between appearance and reality (2011, 2012a, 2018). First, the Platonic model is static and overlooks the temporal, processual nature of human–robot interaction. In reality, what the designer knows as the *program's time*—its internal sequences, pre-set triggers, and control logic—and what the user experiences as *interaction time*—the unfolding rhythm, pacing, and responsiveness—are not two separate realities. They are interwoven components of a single dynamic process in which technical execution and lived experience continually influence each other. Second, this process view shifts focus from opposing reality and illusion to recognizing two interrelated narratives—the designer's and the user's—which evolve together and shape the course of the interaction. Third, Coeckelbergh replaces the reality/illusion framework with that of performance, framing the encounter as co-performance in which users are active participants in constructing meaning, not passive recipients of preprogrammed output. Ultimately, he suggests using phenomenology to understand and go beyond this platonic metaphysics, and this leads to my positive proposal.

4 Phenomenological Alternative

From a phenomenological perspective, reality is not a static, external fact but emerges from the dynamic interplay between a subject and the world. This happens through the body, which acts as a bridge between our inner experience and our environment. A core idea in this perspective is intentionality, which means our conscious experience is always directed toward something outside of ourselves. This outward-directed quality is fundamentally tied to the body. The body provides us with motor intentionality, a non-conscious, skillful way of engaging with our surroundings. For example, when you instinctively know how to reach for a cup without deliberately thinking about it, that's motor intentionality at work. In this reciprocal exchange, the world is actively disclosed to the body as a field of practical opportunities. Objects are manifested as affordances—direct invitations to act that are immediately perceived based on the body's capabilities. For instance a handle is structured for grasping, and a horizontal stretch of ground is available for walking. The world is experienced not as a collection of indifferent things, but as a set of features that physically complement the body's repertoire of skills. This creates a continuous, reciprocal connection between the body and the world, which Merleau-Ponty (2005) called an intentional arc. Our body is directed toward the world, and the world responds by offering possibilities for embodied actions.

This embodied engagement gives rise to a relational conception of reality. This view distinguishes between an object's categorical properties (its fixed, viewpoint-independent features, like a plate's circular shape) and its perspectival properties (how it appears from a specific, embodied viewpoint, like a plate appearing elliptical when seen from

an angle) (Noë, 2004). A key point here is that the elliptical appearance isn't an illusion or a mistake. Perception, therefore, is not merely the passive registration of fixed, factual properties; it is the active grasp of how appearances vary as we move and relate to our surroundings. To perceive an object adequately is to be attuned not only to what it is, but to how it manifests across multiple perspectives. In this sense, perception is a skillful, embodied activity through which reality is enacted.

This approach offers a powerful new way to understand human-artificial agent relationships. Instead of asking whether our emotional experiences with a robot are "real" or "fake," this perspective suggests we should focus on the dynamic patterns of our embodied interactions. The reality of the relationship takes shape within these real-time, reciprocal engagements.

4.1 Human–AA Relations: The Structure of Interaction

In their study of the social robot Pepper, Ujike et al. (2019) observed varied interactions—conversation, eye contact, motivational cues, gestures, and more. A phenomenological approach to human-robot interaction suggests that to truly understand the dynamics at play, we need to shift our focus from the "what-ness" to the "how-ness" of the interaction. Instead of simply cataloging the content of the interactions—like what was said or what actions were performed—this perspective emphasizes the underlying processes and patterns. The most fundamental of these patterns is the action–response cycle, a continuous feedback loop where both the human and the AA are active participants. The cycle begins with one party initiating an action, and the other responding. This response then becomes a new action that the first party reacts to. In the case of the social robot Pepper, it would start with a greeting and a gesture. The patient's initial, sometimes non-verbal, responses—like gazing or a slight smile—are then met by Pepper's adjusted actions, such as verbal praise or new instructions. This ongoing cycle, where each participant's actions are influenced by the other's responses, creates a dynamic, relational interaction. This mutual adjustment produces a smooth, adaptive flow—what phenomenologists call *harmonious interaction*.

Harmonious interaction is crucial to the most fundamental level of experiencing another being as an other (a minded creature)—the *that*-level (Zahavi, 2014)—where we grasp an entity as minded before determining its specific states (*what*) or reasons (*why*). Safdari (2024; 2021) argues that when an action–response loop is harmonious, our anticipations are continually fulfilled, and the entity ceases to be a mere object, becoming an *other*. In the case of artificial agents, this shift produces what he calls an *otheroid*: a non-human entity that, through embodied and reciprocal engagement, is experientially constituted as an *other*. In Pepper's case, this structure fosters an empathic relation in which the robot is no longer perceived as an inanimate tool but as an *otheroid*.

5 Conclusion

In this paper, I have proposed that in order to fully understand anthropomorphic behavior toward AAs, we must take these behaviors seriously i.e. we should not simply dismiss them as mere deceptions or cognitive illusions. To achieve this, I suggest that we turn to the concept of empathy. Accordingly, such behaviors are not misguided or erroneous projections of human traits onto machines, but rather constructive empathic relations. This shift in perspective can help us overcome a long-standing confusion within the literature on anthropomorphism.

The source of this confusion lies in a discrepancy: people tend to ascribe human-like traits to AAs during real-time, embodied interaction, yet resist or even reject such ascriptions when reflecting upon AAs in a detached, reflective manner. According to the framework I propose, this apparent contradiction is not a flaw or inconsistency in human cognition, but rather a constitutive and meaningful aspect of our evolving relationship with AAs.

From this perspective, the empathic relations we establish with AAs in pre-reflective, embodied encounters elevate them from mere objects to what I call otheroids—entities that occupy a relational space between objects and persons. Our interactions with otheroids are characterized by a dual structure: on the one hand, there is a spontaneous, pre-reflective empathic engagement; on the other hand, when we shift into a reflective and abstract mode, this empathic connection becomes unavailable or is actively suppressed. Far from being a problem to be solved, this duality opens up a new way of understanding our relationship with AAs—not as a deficient imitation of human-human relations, but as the emergence of a distinct and positive relational category. The concept of otheroids thus provides a framework through which we can explore and articulate the unique forms of sociality made possible by our encounters with artificial agents.

References

1. Banks, J. (2020). Theory of Mind in Social Robots: Replication of Five Established Human Tests. *International Journal of Social Robotics*, 12(2), 403–414. <https://doi.org/10.1007/s12369-019-00588-x>
2. Breazeal, C. L. (2002). *Designing Sociable Robots*. MIT Press.
3. Coeckelbergh, M. (2010). Robot rights? Towards a social-relational justification of moral consideration. *Ethics and Information Technology*, 12(3), 209–221. <https://doi.org/10.1007/s10676-010-9235-5>
4. Coeckelbergh, M. (2011). You, robot: on the linguistic construction of artificial others. *AI & SOCIETY*, 26(1), 61–69. <https://doi.org/10.1007/s00146-010-0289-z>
5. Coeckelbergh, M. (2012a). Are Emotional Robots Deceptive? *IEEE Transactions on Affective Computing*, 3(4), 388–393. <https://doi.org/10.1109/T-AFFC.2011.29>
6. Coeckelbergh, M. (2012b). *Growing Moral Relations Critique of Moral Status Ascription*. Palgrave Macmillan US.
7. Coeckelbergh, M. (2018). How to describe and evaluate “deception” phenomena:

- recasting the metaphysics, ethics, and politics of ICTs in terms of magic and performance and taking a relational and narrative turn. *Ethics and Information Technology*, 20(2), 71–85. <https://doi.org/10.1007/s10676-017-9441-5>
8. Coeckelbergh, M. (2022). Three Responses to Anthropomorphism in Social Robotics: Towards a Critical, Relational, and Hermeneutic Approach. *International Journal of Social Robotics*, 14(10), 2049–2061. <https://doi.org/10.1007/s12369-021-00770-0>
 9. Damholdt, M. F., Quick, O. S., Seibt, J., Vestergaard, C., & Hansen, M. (2023). A Scoping Review of HRI Research on ‘Anthropomorphism’: Contributions to the Method Debate in HRI. *International Journal of Social Robotics*, 15(7), 1203–1226. <https://doi.org/10.1007/s12369-023-01014-z>
 10. Darling, K. (2021). *The New Breed How to Think About Robots*. Penguin.
 11. Epley, N., Waytz, A., & Cacioppo, J. T. (2007). On seeing human: A three-factor theory of anthropomorphism. *Psychological Review*, 114(4), 864–886. <https://doi.org/10.1037/0033-295X.114.4.864>
 12. Fussell, S. R., Kiesler, S., Setlock, L. D., & Yew, V. (2008). How people anthropomorphize robots. *Proceedings of the 3rd International Conference on Human Robot Interaction - HRI '08*, 145. <https://doi.org/10.1145/1349822.1349842>
 13. Gellers, J. C. (2021). *Rights for Robots Artificial Intelligence, Animal and Environmental Law*. Routledge.
 14. Gunkel, D. J. (2012). *The Machine Question Critical Perspectives on AI, Robots, and Ethics*. MIT Press.
 15. Gunkel, D. J. (2023). *Person, Thing, Robot A Moral and Legal Ontology for the 21st Century and Beyond*. MIT Press.
 16. Hung, L., Gregorio, M., Mann, J., Wallsworth, C., Horne, N., Berndt, A., Liu, C., Woldum, E., Au-Yeung, A., & Chaudhury, H. (2021). Exploring the perceptions of people with dementia about the social robot PARO in a hospital setting. *Dementia*, 20(2), 485–504. <https://doi.org/10.1177/1471301219894141>
 17. Jones, R. A. (2013). Relationalism through Social Robotics. *Journal for the Theory of Social Behaviour*, 43(4), 405–424. <https://doi.org/10.1111/jtsb.12016>
 18. Latour, B. (2015). *Facing Gaia: Eight Lectures on the New Climatic Regime* (C. Porter (trans.)). Polity Press.
 19. Matthias, A. (2015). Robot Lies in Health Care: When Is Deception Morally Permissible? *Kennedy Institute of Ethics Journal*, 25(2), 169–162. <https://doi.org/10.1353/ken.2015.0007>
 20. Merleau-Ponty, M. (2005). *Phenomenology of Perception* (C. Smith (trans.); Taylor and). Routledge. <https://doi.org/10.4324/9780203994610>
 21. Noë, A. (2004). *Action in Perception*. The MIT Press.
 22. Odekerken-Schröder, G., Mele, C., Russo-Spina, T., Mahr, D., & Ruggiero, A. (2020). Mitigating loneliness with companion robots in the COVID-19 pandemic and beyond: an integrative framework and research agenda. *Journal of Service Management*, 31(6), 1149–1162. <https://doi.org/10.1108/JOSM-05-2020-0148>
 23. Placani, A. (2024). Anthropomorphism in AI: hype and fallacy. *AI and Ethics*, 4(3), 691–698. <https://doi.org/10.1007/s43681-024-00419-4>
 24. Puzio, A. (2024). Not Relational Enough? Towards an Eco-Relational Approach in Robot Ethics. *Philosophy & Technology*, 37(2), 45. <https://doi.org/10.1007/s13347-024-00000-w>

Otheroids or Anthropomorphism?

- 024-00730-2
25. Safdari, A. (2024). Toward an empathy-based trust in human-otheroid relations. *AI & SOCIETY*. <https://doi.org/10.1007/s00146-024-02155-z>
 26. Safdari Sharabiani, A. (2021). Genuine empathy with inanimate objects. *Phenomenology and the Cognitive Sciences*. <https://doi.org/10.1007/s11097-020-09715-w>
 27. Sharkey, A., & Sharkey, N. (2012). Granny and the robots: ethical issues in robot care for the elderly. *Ethics and Information Technology*, 14(1), 27–40. <https://doi.org/10.1007/s10676-010-9234-6>
 28. Sharkey, A., & Sharkey, N. (2021). We need to talk about deception in social robotics! *Ethics and Information Technology*, 23(3), 309–316. <https://doi.org/10.1007/s10676-020-09573-9>
 29. Sparrow, R., & Sparrow, L. (2006). In the hands of machines? The future of aged care. *Minds and Machines*, 16(2), 141–161. <https://doi.org/10.1007/s11023-006-9030-6>
 30. Thellman, S., de Graaf, M., & Ziemke, T. (2022). Mental State Attribution to Robots: A Systematic Review of Conceptions, Methods, and Findings. *ACM Transactions on Human-Robot Interaction*, 11(4), 1–51. <https://doi.org/10.1145/3526112>
 31. Turkle, S. (2011). *Alone Together: Why We Expect More from Technology and Less from Each Other*. Basic Books.
 32. Ujike, S., Yasuhara, Y., Osaka, K., Sato, M., Catangui, E., Edo, S., Takigawa, E., Mifune, Y., Tanioka, T., & Mifune, K. (2019). Encounter of Pepper-CPGE for the elderly and patients with schizophrenia: an innovative strategy to improve patient's recreation, rehabilitation, and communication. *The Journal of Medical Investigation*, 66(1.2), 50–53. <https://doi.org/10.2152/jmi.66.50>
 33. Vertesi, J. (2015). *Seeing Like a Rover: How Robots, Teams, and Images Craft Knowledge of Mars*. The University of Chicago Press.
 34. Winkle, K., Caleb-Solly, P., Leonards, U., Turton, A., & Bremner, P. (2021). Assessing and Addressing Ethical Risk from Anthropomorphism and Deception in Socially Assistive Robots. *Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*, 101–109. <https://doi.org/10.1145/3434073.3444666>
 35. Zahavi, D. (2014). Self and Other: Exploring Subjectivity, Empathy, and Shame. In *Oxford University Press*. <https://doi.org/10.1017/CBO9781107415324.004>

Bestowing Legal Personality on AI: A South African Perspective

Kristi Erasmus^{1[0009-0006-1488-0194]}

[kristierasmus@gmail.com]

¹ STADIO Higher Education, Cape Town, South Africa

Abstract.

Granting legal status to Artificial intelligence (AI) could address much of the uncertainty that currently surrounds accountability and liability of AI systems. As AI lacks sentience and expression of its own will, it cannot be deemed a natural person however juristic or legal personality may possibly be imposed on it. Consideration of the South African common law requirements for juristic personality suggest that AI as a non-human entity meets the prerequisites and consequently should be bestowed with juristic personality in the same way it is bestowed on companies and similar corporations. The research argues that AI meets the South African common law requirements for bestowing juristic personality on non-human entities and postulates that on this basis the South African Companies Act may serve as the primary guiding legislation on which basis a specific AI statute and associated regulations may be drafted and entrenched. By means of a doctrinal desktop study this research considers arguments for and against bestowing juristic personality on AI systems and argues in favour of juristic personality being bestowed on AI discussing the potential benefits that may arise from same. The research by no means attempts to provide a definitive answer to the questions posed but seeks to stimulate academic discussions and contribute to the existing body of knowledge exploring the legal status of AI from a uniquely South African perspective.

Keywords: Artificial Intelligence, Juristic Person, Companies Act 71 of 2008.

1 Introduction

Artificial Intelligence (AI) has infiltrated every aspect of personal and commercial life [1]. It has evolved to become a necessary commodity in business transactions and activities, required to increase productivity, profit and competitive market share [2]. 78% of companies worldwide have confirmed their use of artificial intelligence in at least one business function, up 6% from what was reported during 2024 and up 23% from what was reported in 2023 [3]. In South Africa the most recent study by World Wide Worx has confirmed that 90% of participating respondents currently use or intend to use generative AI in their business operations, confirming that the adoption of generative AI in South Africa is a catalyst for innovation and growth [4]. However, despite the efficiency and productivity that AI provides in a business environment, it remains significantly unregulated in South Africa. To date no legislative provisions regulating

AI has been promulgated in South Africa. Only a policy framework [5] and an AI National Government Summit Discussion Document [6] have been published and a process to obtain inputs for a national AI plan have been launched.[7]

The limited regulation of AI is problematic given the uncertainty that arises in respect of ensuring its proper development and deployment, allocating creative ownership and copyright where it generates novel literary, artistic or musical works and imposing liability and accountability where harm should be caused through or by it. The latter is of primary concern given the tendency of generative AI systems to hallucinate and provide false information, on which reliance could have severe negative consequences resulting in monetary or reputational loss, harm or damage. The far-reaching negative consequences of AI has recently been evidenced in three reported South African court judgments where three attorney firms, on separate and unrelated instances, blindly relied on an AI system to draft court documents without fact checking the AI system output, which contained false law precedents [8]. This resulted in the relevant firms being penalised for misleading the court. This not only holds reputational damage for the relevant attorney firms but also translates into additional monetary damage suffered by the client given wasted and consequent delayed legal proceedings. Similar harmful consequences of AI have been witnessed in the USA where legal briefs with fake citations have also been seen along with the AI system CHATGPT wrongfully accusing a professor of sexual assault [9].

It is clear that as AI is increasingly affecting our lives, our legal systems are late in providing appropriate regulations either through amendment of existing laws and legal systems or in entrenching new laws and systems to appropriately govern and regulate AI [10]. The obstructive uncertainty, specifically in respect of liability and accountability for loss, harm or damage caused by AI, could possibly be addressed by assigning legal status to AI necessitating only a slight amendment of our existing laws.

As AI refers to intelligence demonstrated by machines, distinct from the intelligence demonstrated by humans and animals which involves consciousness, self-awareness and emotionality [11], AI cannot be deemed a natural person as only humans can be assigned such title. However, the question then arises as to whether we may assign juristic personality to AI and what effect or consequence this may have.

Consideration of the South African common law requirements for juristic personality suggest that AI as a non-human entity meets the prerequisites and consequently should be bestowed with juristic personality in the same manner in which it is bestowed on companies and similar corporations. This in turn begs the question that if AI is deemed a juristic person, whether the existing Companies Act 71 of 2008 may serve as a guiding legal framework for the drafting of an appropriate statute regulating AI in South Africa.

A definition of AI will firstly be introduced to demonstrate what is meant by AI in today's terms differentiating realistic AI from science fiction. This will be followed by

a differentiation between natural and juristic personality, with confirmation of what the consequences and effects of juristic personality is and on what non-human entities it can and has been bestowed on. Hereafter the South African common law requirements for bestowing juristic personality will be discussed, with arguments in favour of bestowing AI with juristic personality being set out. This will be followed by considering the extent to which the South African Companies Act 71 of 2008 may serve as a guiding legal framework for the drafting of a uniquely South African AI statute.

2 Research Methodology

The research paper adopts a qualitative, doctrinal desktop study. Through in-depth review of various scholarly articles, research papers, conference papers and existing guidelines, common patterns in arguments presented for and against granting juristic (legal) personality to AI have been identified and are set out herein. Whereafter through a process of deductive reasoning based on the review of the sources as described and consideration of common law precedents and theories on juristic personality, the research question is sought to be addressed and future research possibilities confirmed.

Limitations of the research must be acknowledged. Given the significant pace at which AI is continuously being developed and deployed, the research undertaken now and arguments presented herein may be obsolete within a couple of years or even months after conclusion hereof. Furthermore, the lack of a universal definition of AI and of what ‘personhood’ entails given different foreign jurisdictions and legal systems also limit the extent to which the arguments may be deemed applicable or acceptable to specific jurisdictions and legal systems. It is additionally noted that the challenges, risks and liabilities that may arise from AI use and development are extensive and far reaching. However, discussion and consideration of all nuances associated and arising from AI use and development is beyond the scope of this research and consequently only specific aspects related to the accountability and liabilities arising from AI are discussed herein.

Consequently, this research by no means attempts to provide an all-encompassing discussion on AI or provide a definitive answer to the questions posed but simply seeks to stimulate academic discussions and contribute to the existing body of knowledge exploring the legal status of AI from a uniquely South African perspective.

3 Defining AI

Despite reasonable and practical explanations advocating for a universally accepted definition of AI, there is no uniform, globally accepted definition of AI [12][13]. AI is generally accepted to refer to machines or computers that display cognitive abilities required to perform tasks and activities that generally require human intelligence including but not limited to learning, problem solving, perception, language understanding, decision making, logic, thought and reason [14].

AI systems can be divided into three broad categories dependent on their capacities and level of autonomy. The AI we encounter today is generally known as weak AI, which are AI systems designed to perform specific tasks within a specified domain. These systems do not possess general intelligence or the ability to perform tasks or activities beyond what they are programmed to do [14]. General AI, also known as strong AI, is currently only hypothetical and refers to systems that possess human level intelligence and the ability to learn and perform various tasks across a range of different domains [14]. Similarly, Autonomous AI, currently only a future notion, are AI systems capable of operating completely independently without human oversight or intervention, making decisions and taking actions without human control or interference [14]. Currently strong and autonomous AI remain hypotheticals while weak AI is the reality of AI we see today. AI is not capable of independent thought or consciousness and is limited to performing specific tasks within specified domains as per its programming.

It should be noted that the term AI is often used incorrectly. Contrary to the manner in which reference is made to AI, AI is not itself a computer, machine or agent but is the term assigned to various ‘intelligent’ systems, that is processes embodied within technology. AI is an umbrella term the same way in which ‘fruit’ is the umbrella term for various fruits such as apple, banana, orange, pear, pineapple [15][16]. Consequently, AI encompasses a diverse set of approaches and techniques made up of algorithms that enable machines, computers and similar technology to perform tasks that traditionally require human cognitive ability and intelligence [17]. AI, a branch of computer science, is in simple terms various algorithms, which are instructions and rules formulated by programmers, that enable a computer (or robot controlled by a computer) to mimic human intelligence.

Reference to AI today generally refers to machine learning [17]. Simply, machine learning refers to layers of algorithmic instructions according to which vast quantities of data and inputs are processed, identified and sorted, extracting structured data from unstructured data. Based on data that is incorporated, algorithmic formulations of the AI are effectively trained on identifying patterns and commonalities in the data provided and from which, based on statistics, it is then able to make predictions and identify the next pattern or word. This allows it to predict an output that is statistically correct, mimicking the thinking of a natural human being and creating the impression that it possesses cognitive abilities [18].

Machine learning is often used interchangeably with deep learning. However, deep learning is a sub-category of machine learning, with the “deep” of deep learning referring ultimately to the extensive layers of interconnected neural networks which make up machine learning and through which algorithms are trained. The deeper the deep learning the less human involvement is required and better the ability of the AI to itself acquire its own input data and sort same.

However, a computer or machine cannot understand natural human language. Computers and machines communicate using artificially created, specialised formal languages, that are highly constrained, mathematically structured algorithms, such as Python or Javascript [19]. The languages are created specifically for unambiguous mechanical interpretation and thus are very limited in what they can express or convey in terms of the complexity and nuance of real word ideas or phenomena in comparison to natural human languages [19].

Natural language processing, an interdisciplinary field of study combining linguistics with computer science, and a subfield of AI, has provided computers and machines with the ability to ‘understand’ spoken and written human language by treating language as data and identifying statistical matches to specific patterns, [19] translating a limited range of formal programming instructions into known and determinate computational actions and executes them [19]. Nevertheless, a significant limitation of natural language processing is its inability to grasp the underlying, ambiguous meaning of words [19].

This is where large language models, a subfield of natural language processing, becomes important. Large language models are AI systems designed to understand and generate human language [19]. Large language models, a subset of natural language processing, consists of neural networks that are trained to understand and predict human language by learning language patterns and structures [20][21]. Given the layers of neural networks, it is also deemed to be a subset of machine learning and consequently also a subset of deep learning [22][23]. The large language model uses copious amounts of text data as its input and on which basis it statistically determines the most likely word to follow another word enabling it to accurately predict what word would follow [22][23]. In its most basic form, it is a glorified word predictor. With large language models, AI systems are able to grasp the syntactical variation, semantic ambiguity and abstract complexity that is typically part of natural human language [19]. Additionally through machine learning it is able to improve its accuracy and eventually predict not only the next word that would follow but what sentence, paragraph or full content would follow, thereby generating new output similar but different to the input data, in natural language writing, thereby generating content that looks like human generated content [22][23][24][25].

Large language models form the operational basis of generative AI’s. Generative AI, also known as Gen AI [20], is an umbrella term of AI systems that are associated with human creativity with the term ‘generative’ serving to demarcate the purpose of the system as being focused on human creativity [19]. Generative AI makes use of a combination of machine learning, deep learning and its neural networks together with natural language processing and its large language models to generate novel content from the vast quantity of unlabeled data, and some labelled data, on which it has been trained based on a prompt or request provided by a user [26]. The output serves as the answer to the user prompt or request and may be anything from legal document to lyrics to a song to a recipe based on the contents of a fridge [27].

Large language models can effectively create and understand complex written language texts such as research papers, fictional stories, reasoning puzzles and newspapers [19]. Apart from its text analysis abilities it is also capable of synthesizing and reasoning about facts and the physical world, problem solve with abstract ideas and notions and work with modalities beyond text including images, sound and video [19]. This is then combined with natural language processing, which allows a computer to understand text and spoken words as a natural person would [28][29].

Consequently, reference to AI within this research should be understood as reference to large language models, as narrow AI that through deep learning, natural language processing and layers of neural networks generate and understand novel content, mimicking human intelligence, appearing to think critically, problem solve, make decisions, rationalize, learn, and plan.

4 Personality Types – Natural v Juristic Personality

Most western legal systems differentiate between natural personality and juristic personality which is said to be encompassed within legal personality. Legal personality is explained by Smith (1928 Yale Law Journal) as follows: “to be a legal person is to be subject of rights and duties. To confer legal rights or to impose legal duties, therefore, is to confer legal personality” [30][31]. Natural personality is bestowed on natural persons, that is all human beings, automatically from birth until death and is focused on the protection of fundamental human rights [30][31]. Juristic personality is bestowed in terms of statute on non-human entities such as companies and similar association of persons; from the moment their incorporation is registered in terms of law until they are deregistered. Juristic personality is imposed for practicality and legal security onto entities such as businesses, organisations or corporations which have been granted title of “legal person” artificially by humans to allow those entities to exist and exercise their own rights and obligations [30][31].

The question on whether juristic personality should be bestowed on AI in a highly contested debate that has a clear divide between those in favour and those against it [32]. One school of thought argues that AI will eventually reach a point where it surpasses human intelligence and consequently should be granted rights and obligations like what is granted to companies as juristic persons [10]. On the opposing side it is argued that as AI is the outcome of human programming, AI lacks autonomy and free will and consequently it should not be recognised as a juristic person [10].

Reasons for granting juristic personality to non-human entities such as companies and similar corporations include providing them with separate legal existence to enable them to hold rights and duties in their own names. Additionally, by affording juristic personality to companies and corporations it permits people to engage directly or contract with the company as an establishment as opposed to the various individuals

associated with it. This establishes a direct line of contact, making it easier for liability to be imposed where fault should arise. Furthermore, bestowing juristic personality on companies ensures better and more efficient facilitation of transactions, managing assets and debts, bookkeeping, tax duties and administration of bank accounts. Also, by bestowing juristic personality on companies a separate individual and independent entity is introduced that will be directly liable where loss or damage should be caused. Moreover, bestowing juristic personality on companies ensure there is no suspension or disruption in economic activities due to a change in the individuals associated with it such as may arise by reason of death or resignation, which in turn provides for greater trust and legal security in the market and in society [31].

The arguments set out above may serve as valid arguments raised in favour of bestowing juristic personality on AI systems. Reasons for granting juristic personality to AI, a non-human entity, includes providing them with separate legal existence to enable them to hold rights and duties in their own names. Additionally, by affording juristic personality to AI, it permits people to engage directly or contract with the AI itself as a person as opposed to the various individuals associated with it. This establishes a direct line of contact, making it easier for liability to be imposed where fault should arise. Furthermore, bestowing juristic personality on AI may further ensure better and more efficient facilitation of AI transactions, management of AI's assets and debts, bookkeeping, tax duties and administration of its bank accounts. Also, by bestowing juristic personality on AI a separate individual and independent entity is introduced that will be directly liable where loss or damage should be caused. Moreover, bestowing juristic personality on AI may serve to ensure there is no suspension or disruption in economic activities due to a change in the individuals associated with it, which in turn provides for greater trust and legal security in the market and in society [31].

It should be noted that bestowing legal personality on non-human entities is not unheard of and has in recent years increasingly been recognised in different legal systems. Apart from natural and juristic personality, the concept of environmental personality has also recently been recognised in some Western legal systems. Environmental personality is like juristic personality, in the sense that it entails bestowing legal personality on non-human entities, that is elements of nature. This is done so that certain rights may be attributed to the natural entities so as to enable persons and advocacy groups to act on their behalf in their preservation and safe keeping without incurring problems regarding legal standing, bestowing legal rights and executing legal duties and obligations [31]. This has been seen in New Zealand where legal personality has been bestowed on a national park, Te Urewera, and a river, Whanganui, through legislation. Similarly in USA, Lake Erie has been bestowed with legal personality to enable residents to institute legal action against polluters on the lake's behalf [31]. While legal personality has been conferred on temples and the Ganges and Yamuna rivers [33] in India and an entire ecosystem in Ecuador [34].

The so called "environmental personality" serves to counter the power balance between governments and nature, permitting principles and provisions of private law to

apply to nature as opposed to only administrative environmental law alone being applicable or relied on to preserve these natural entities. This means, as has been seen with Lake Erie in USA, that through applicable private laws, residents, non-governmental agencies and advocacy groups are permitted to institute legal action in the name of the specific natural entity to protect it against dangerous or pollutive activities. The natural entities, bestowed with legal personality, are recognised as separate independent entities that are afforded rights such as protection and preservation. However, liability, obligations and duties cannot be bequeathed to these entities alone and effectively require human agents to discharge duties imposed on such natural entities [13][31]. The same is required with juristic personality of companies where company rights are enforced and duties are executed on the company's behalf by a board of natural human directors.

Consequently, the personality granted to the natural entities serves as a pretence to address problems over legal standing and ability to institute legal action. By affording personality to certain natural entities, it enables human individuals to act on behalf of the non-human rights holder rather than requiring it to establish standing in its own capacity [34]. Granting juristic personality to AI would have the same effect but in reverse. It would allow individuals to hold AI liable and accountable for harm caused by or through its use by holding AI to certain specific legal duties deemed executable by the individuals associated with it, without having to prove legal standing and capacity.

5 Arguments for imposing juristic personality on AI

There are several reasons why it would be beneficial to extend juristic personality to AI.

Firstly, by bestowing juristic personality on AI it creates the opportunity for establishing a legal framework in terms of which AI can be held liable for harm or damage caused by it [14]. Currently there is a lack of accountability attributable to a person in so far as AI is concerned as, determining who had control over the outcome of AI becomes increasingly difficult the more autonomous the AI system becomes. In terms of the Law of Delict liability is imposed for harm or damage arising from fault or negligence, however the more autonomous AI systems become the harder it is to foresee what the outcomes of their actions may be, making it more difficult to determine what would constitute fault in such circumstances. Where a person cannot predict harm or had no control over conduct causing harm, liability cannot be imposed unless strict liability applies. Whether strict liability may apply to the developers and manufacturers of AI is currently debated in academia but unfortunately discussions on same is beyond the scope of this research. Bestowing juristic personality on AI will ensure that end users who may suffer harm, loss or damage from the AI use would have a specific entity from which damages could be claimed in the same manner as is done with Companies as a juristic person today [35].

Secondly if juristic personality is bestowed onto AI, it will allow AI to directly participate in legal transactions such as contract conclusion and property ownership. The AI will itself be a named party to the contract or transaction, giving rise to rights and duties that vest in the AI itself. This will provide for the streamlining of legal transactions as it reduces ambiguity as to who should bear rights and obligations. The parties directly affected by the transaction will be limited to the named entities and not subject to individuals entering transactions or concluding contracts in their own name but for the benefit of another. Additionally, should juristic personality be bestowed on AI the AI itself may enter contracts and transactions reducing the extend of human intervention required. Both aspects will lead to greater efficiency and economic growth [14]. Granting AI juristic personality would streamline legal transactions because it provides a single legal identity for AI, ensuring contracts, liability, and compliance can be attributed directly to the AI itself as a named party in the contract or transaction.

Thirdly it will permit AI to own intellectual property in terms of the content it generates which could include artwork, designs, patents and other financially valuable outputs [31]. The latter may serve as the patrimony from which claims for loss or compensation could be paid, further addressing the liability gap that currently exists with AI.

Fourthly, by affording AI legal status, AI itself will be incorporated into existing legal frameworks, which enables existing legal frameworks to directly address emerging issues and challenges associated with AI use as and when they arise, countering the reactive as opposed to proactive nature of legislation and law making [14]. Additionally it should be kept in mind that as we have an existing legal frame-works (statutes, regulations and case law) across various branches of law that regulate juristic persons and juristic personality, should AI be deemed endowed with juristic personality, flexibility is provided in our existing legal systems in terms of which each field of law will have the freedom to assess the legal issues posed by AI within its own boundaries , subject to its own specific statutes and regulations, requiring only slight amendments as opposed to entrenchment of completely new laws [35].

Lastly, as investors are risk adverse, they are more likely to invest capital if there is certainty as to how AI will be regulated, who will bear liability and how disputes will be resolved. The incorporation of AI into our current legal systems will provide greater certainty, providing predictability boosting investor confidence and ultimately encouraging innovation and economic growth. Consequently, the legal recognition of AI could provide a foundation for increased investment and collaboration which in turn could lead to creation of new industries and markets centred around AI, promoting economic growth and employment opportunities [14].

Consequently, if juristic personality is bestowed on AI, AI would be deemed separate and independent person, possessing the ability to hold its own rights and duties, own its own assets and revenue and be liable for its own debts and liabilities. Being deemed to be an independent ‘person’ in its own right, means AI would be deemed to have the capacity to enter contracts in its own name and hold rights and duties separate

from its owners, developers, manufacturers or users. This would enable AI to enter contracts in terms of which it would acquire property and associated assets, which in turn would represent the proprietary pool from which damages and compensation for harm or loss caused could be paid by the AI itself.

6 Arguments against imposing juristic personality on AI

However, despite arguments in favour of bestowing juristic personality on AI, AI is deemed perilous to human beings given its role and functionality in surveillance, profiling, manipulation, job losses, automatization of important decisions regarding health and well-being and potential breach of privacy [31]. Consequently, most scholars reject the idea of granting personality to AI as granting personality is deemed to be dangerous and contrary to human interest [31].

Granting juristic personality to AI raises ethical concerns in so far as AI rights may conflict with human rights. For example, property rights and rights to participate and enter legal transactions could lead to situations where AI effectively competes with humans for land, water, power and resources alike, resulting in potential harm to human beings [14].

Additionally, it is said that big technology and AI companies could evade liability by transferring risks and liability to AI systems itself were AI systems to be bestowed with juristic personality [10]. It is argued that bestowing juristic personality onto AI would present a means by which manufacturers, developers and producers of AI may externalize risk and evade liability by holding AI directly accountable for loss and damage [31]. The same way the juristic personality is abused where companies are used to facilitate illegal activities in the company's name and not the names of shareholders or directors. This would ultimately serve to weaken the market incentives to comply with safety standards and regulations [31]. Opponents argue that it would better serve liability claims and claims for damages by imposing liability and accountability directly on the creators, operators and owners of AI as opposed to the AI itself as a juristic person [14][31]. It is said that imposing responsibility on the human agents behind AI would serve to ensure that liability is appropriately placed on those that have the power to control and influence the AI system itself [14][31].

This argument however disregards the so-called black box effect of AI. Certain characteristics of AI such as complexity, opacity, machine learning and autonomous decision making make it difficult to understand how AI systems work and function in all instances. In some instances, even the AI system developer does not know how a certain outcome or output produced by AI has been achieved. This is known as the black box effect [31]. Due to this black box effect limiting the system developer's knowledge and understanding on how a certain outcome or output produced by AI was achieved, liability cannot be imposed as liability cannot be imposed where the owner has limited

foresight, knowledge or control over the AI conduct. This hinders imposing liability for harm or loss caused and raises potential obstacles in establishing causation [13].

Additionally, AI systems are often developed and designed by way of collaboration between researchers, engineers, programmers and designers making it difficult to identifying a specific individual to whom liability may be attributed as it is difficult to determine whose actions caused the resultant harm that may arise through the AI use. Loss or harm arising from AI may be due to any number of reasons, including but not limited to faulty design, error in coding of AI, defective sensors, incompatible software updates, fault or negligent human oversight, improper use of AI or inappropriate or outdated data for training [31]. Who should bear liability remains uncertain. Given the complexity, opacity, connectivity and autonomy of AI systems it would make it very difficult in practice to trace back a specific harmful action of an AI system to a specific human agent input or decision [35].

Ultimately, end users of AI would be better served where juristic personality is bestowed on AI thereby providing a one on one relationship between the AI system and end user, "...whereby the AI [system] effect that affects them visibly is caused by a specific entity...", being the AI system itself which is bestowed with juristic personality and thus is deemed to be a named party to the transactions [35]. This will ensure that the end user has clarity against whom legal action should be imposed should harm, loss or damage arise. This would enhance liability as the confusion that arises where there is a mass of multinational organizations and stakeholders involved in a transaction and who should bear liability would effectively be avoided. The AI system could itself be held directly liable and legal action would be instituted against the AI itself, in the name of the AI as opposed to a victim having to determine against which individual in a mass of multination organizations and various stakeholders involved in AI manufacturing, development and deployment committed a wrongful or harmful act that caused the harm or loss and in whose name legal action should be instituted.

Apart from the potential conflict of interest and rights between AI and humans and manufacturers, developers and owners evading risk and liability, it is additionally argued that juristic personality should not be imposed on AI as any claims for loss or damage caused by AI would be limited to the value of the AI system itself. It is contended that imposing juristic personality on AI for purposes of establishing a framework by which AI can itself be held liable for loss or damage caused is without merit as damage claims against AI would be limited to the revenue earned by the AI system or to value of the AI system itself. It is argued that this value would be reduced to very little given that the developers, manufacturers and producers would most likely draw profits from the AI as renumeration for their time and effort in developing and making the AI function effectively [31].

Additionally, it is argued that as AI lacks consciousness, emotions and subjective experiences it should not be bestowed with juristic personality as these characteristics are unique to legal personality. It is argued that bestowing legal personality on AI,

which lacks the essential characteristics of persons, would ultimately serve to devalue the concept of legal personality and would be inappropriate [14]. However, it should be kept in mind that human consciousness or will is not a prerequisite for juristic personality as same is not required or deemed a relevant criterion in terms of our laws granting juristic personality to corporations, companies and similar associations of persons [14] who have no ‘soul to damn’.

Lastly it is argued that bestowing legal personality onto AI would serve to introduce complexity into existing legal systems and laws as it would require development of new legal concepts and principles not previously encountered or provided for, which would need to be unique to address the characteristics and capabilities of AI systems. This could lead to confusion and uncertainty for enforcement of existing laws and regulations [13][31]. Consequently, it is argued that new laws should be enacted to cater for the unique characteristics and capabilities of AI as existing legal provisions would not be sufficient to appropriately govern AI.

The latter argument is however in direct conflict with the solutions proposed by existing scholars who argue against granting juristic personality to AI systems directly. The opponents argue that existing laws can and should be amended and extended to address liability, loss and harm caused by AI without bestowing it with juristic personality [31], saying that existing laws can be amended to appropriately govern AI. It is argued that legal personality need not be conferred to AI as promulgating stricter safety requirements and standards and amending or extending existing laws to address and cover AI related damage or harm, would be the better option than bestowing AI with legal personality for the sole purpose of addressing the liability and accountability gap [31]. This is however in direct conflict to the above argument arguing that bestowing juristic personality on AI would lead to too much confusion given our existing legal systems and consequently new laws should be enacted.

Although these arguments are raised against bestowing juristic personality on AI, the same arguments could be raised in opposing the granting of juristic personality to companies and corporations. Companies, as juristic persons, directly compete with humans for land, water, power and resources alike. Similarly, a company’s juristic personality has and is abused and used by its owners to evade liability, externalizing risk for illegal activities that cause harm and damage through its actions. Furthermore, claims for damages against companies are limited to the net value of the company, which could be a negligible amount given that owners and investors will draw from company profits and assets as renumeration and as return on their investments. Lastly it is conceivable that existing laws could be amended and extended to appropriately regulate companies as opposed to bestowing it with juristic personality itself. Even though the same arguments raised against bestowing juristic personality on AI can be raised against bestowing juristic personality on companies and similar corporations, the latter has been recognised and formed part of our legal system for many centuries despite the counter arguments raised [31][36].

From the above it is clear that the arguments against imposing juristic personality on AI contradict each other. Additionally, given the fact that the same arguments against imposing juristic personality on AI could be raised against imposing juristic personality on companies and corporations, it can be contended that arguments against bestowing AI with juristic personality lacks pragmatism and credibility.

7 Juristic Personality: Common Law in South Africa

A juristic person refers to a social entity or association of persons that is deemed to have an independent right of existence, enabling it to be the bearer of judicial capacities such as holding subjective rights and bearing legal obligations in its own right and name [36]. Before a social entity or association of persons can be deemed to be a juristic person, certain formalities as set out in the South African law must be met.

Having regard to these formalities, juristic persons can be divided into three groups: Firstly, organisations that can only be established by permission of the government where the organisations existence and capacity to act as a legal subject is regulated and governed by statute. This will generally include organisations of national interest such as EKSOM which was established as a juristic person in terms of the Electricity Act 41 of 1987 [36].

The second category of juristic persons is managed by government in the interests of society however the executive need not approve its incorporation as with the first category [36]. These associations need only register in terms of the relevant statute. This would include organisations incorporated for purposes of generating profit which must be registered in terms of the Companies Act 71 of 2008. The nature of the organisation will determine under which statute registration of the incorporation must be affected.

The final group of associations classified as juristic persons exist as such not due to government intervention but since they meet the common law requirements for the establishment of a juristic person [36].

The common law requirements include [36]: the entity must continue to exist despite changes in its members and must thus possess perpetual succession. Additionally, the entity should be the bearer of its own rights and duties, separate and distinct from the rights and duties of the individuals behind it and should possess or have the capacity to possess its own property. Lastly the entity should strive toward a legal, predetermined goal.

Legal conditions for juristic personality are provided for purposes of ensuring public order, providing for regulation of the market, safeguarding commercial efficiency, and appropriate allocation of risk, liability and legal security [31]. The same could be achieved were AI to be bestowed with juristic personality.

Accepting that any entity that meets the common law requirements of juristic personality may be bestowed with same through operation of the common law, it can be argued that any entity that has perpetual succession, is the bearer of rights and duties, can own or has the capacity to own property and which strives toward a pre-determined goal may be deemed to be a juristic person.

8 Application Common Law Requirements to AI

An “entity” is defined as “a thing with distinct and independent existence” [37]. AI is, by its very nature, an entity that exists independently and distinctly as a tangible or intangible component of a systems processes and thus is not regarded as being part of any person. Much like a computer, it is an object, a thing, that possesses its own separate and independent existence. AI furthermore possess perpetual succession as its continued existence is not dependent on one specific user or developer but can continue to exist irrespective of changes in users or developers. AI furthermore has recently been recognized as having the capacity to own intellectual property with DABUS, an AI system, being awarded a patent by the South Africa’s Patent Office (SAPO) who operates under the patronages of the Companies and Intellectual Property Commission (CIPC), the regulatory body regulating enforcement and compliance of the South African Companies Act 71 of 2008 [38]. AI is also programmed toward achieving a pre-determined goal, which is to enhance the thinking, creativity and productivity of its users. Consequently, the only requirement that cannot be directly met at present by AI is its ability to bear rights and duties. However, considering AI’s ability to own intellectual property, as we have seen with DABUS, it implies that it can be the bearer of rights and duties and thus all requirements are met for juristic personality to be bestowed on it.

If juristic personality is bestowed on AI, the AI would itself be deemed to be a juristic person separate and independent of its developers and programmers. It would hold its own rights and duties separate from those of its developers and programmers, and it would own its own property and be liable for its own debts and liabilities. Lastly should any wrong be committed against it, it would be for the AI itself to proceed with legal action and not its developers or programmers.

9 AI under the Companies Act – Legislative Analogy

AI, like a company, can be conceptualised as a juristic person constructed from many individual contributions. A company is often likened to a house built of bricks: each shareholder invests capital and, in return, acquires a portion of the “bricks” that make up the structure. These bricks represent shares, which carry with them both financial interests and governance rights. The greater the number of bricks owned, the stronger the shareholder’s voting power and influence over the company’s affairs.

In a similar way, the creation of an AI system may be viewed as the construction of such a “house.” Developers and programmers, by investing resources, expertise, or intellectual capital, can be understood as acquiring “bricks” within the AI system. Their proportional contributions determine the degree of control or decision-making power they exercise over the governance and functioning of the AI system.

This analogy provides a foundation for considering how sections of the Companies Act 71 of 2008 could with appropriate amendments and adaptions serve as a model to regulate AI within south Africa. However, before consideration is given to how principles of corporate law and provisions of the Companies Act may be amended to apply to regulation of AI systems an important conceptual inconsistency between AI systems and companies as juristic persons should be addressed.

Although an AI system may be deemed a juristic person in the same way a company is, it does not naturally have shareholders (security holders) or directors. This creates a conceptual difficulty. However, the conundrum can be addressed by assigning the traditional corporate roles of shareholders and directors to those most directly involved in the creation and operation of AI systems, that is developers and programmers.

Just as shareholders provide capital and conceptualise the purpose and structure of a company, developers contribute intellectual and financial resources in designing and constructing the AI system. Their role in initiating and shaping the AI system parallels the way shareholders provide the foundational vision and investment for a company.

Similarly, programmers, following the instructions and objectives set by developers, direct the AI system’s functioning. They encode the necessary algorithms and operational parameters, ensuring that the AI system operates as intended and in alignment with the developers’ objectives. Thereby fulfilling the role of directors who in companies act on behalf of shareholders, exercising oversight and directing the day-to-day management of the company in the best interests of shareholders.

Thus, while AI systems lack shareholders and directors in the conventional sense, the functions these roles perform in corporate law can be transferred to developers and programmers. This conceptual adaptation provides a workable governance model for AI juristic persons, preserving accountability and oversight in a manner consistent with established principles of corporate law.

As the AI system would like a company be a fictional, hypothetical person with no soul to damn and no body to kick [34], the AI developers and programmers would like company directors and shareholders be required to act on behalf of the AI in so far as enforcement of rights and execution of duties are concerned. Consequently, should legal action be instituted by or against the AI, the developers and programmers would in the same manner as company shareholders and directors proceed with the legal action in the company’s name proceed with legal action in the AI system’s name and on the AI system’s behalf.

This establishes a relationship suggestive of a holding and subsidiary relationship. It should be noted that the owners, developers, and programmers of an AI system typically operate through a company, which is itself a juristic person, such as Open AI who owns Chat GPT and Microsoft who owns Co-Pilot. As a result, one juristic person, the company comprising the developers, and programmers would effectively own the AI system, which is also considered a separate juristic person.

The AI system, if deemed to be a juristic person, would as provided currently in section 3 of the Companies Act (71 of 2008), be the subsidiary of another juristic person, that is the company to which the programmers and developers belong. This could be called the AI holding company. This latter company, as its own juristic person, would own the AI system as it would be able to directly or indirectly exercise or control the exercise of majority of the voting rights associated with the issued securities of the AI system or would have the right to appoint or elect or control the appointment or election of programmers, assigned the traditional role of directors in corporate law, of that AI system where the programmers come together for a meeting to vote on a matter related to the AI system.

Questions may arise on how liability and accountability could appropriately be imposed if AI systems are deemed to be juristic persons given that the individuals associated with a juristic person, such as shareholders and directors, enjoy limited liability. Consequently, the question arises how accountability and liability could be imposed on an AI system when its owners, developers and programmers enjoy limited liability by virtue of the fact that the AI system is itself deemed to be its own individual, separate and independent person liable for its own debts and liabilities, for which owners cannot be held liable. Argument could be made that where the AI system should cause harm or damage, liability would be imposed on it, and it would need to compensate for the loss or damage caused from its pool of property which it owns given the right of juristic persons to own rights and property. Should the pool of property be insufficient, the AI holding company, being the owning company, such as OpenAI or Microsoft, may itself be held liable. Should same still not yield monetary compensation, piercing of the corporate veil as permitted in the case of holding and subsidiary companies could be resorted to. Piercing of the corporate veil would mean that the juristic personality of the group of companies, and separate existence afforded to the holding and subsidiary company is disregarded. Consequently, the owners, developers and programmers operating through a juristic person themselves would be held personally liable, losing the statutory limited liability protection afforded by juristic personality.

In so far as AI users are concerned, it should be accepted that AI users, referring to students, scholars and business professionals who use AI in their daily life , research or business activities would be clients and customers of an AI system the same as company clients and customers are the users of the company services it renders as its primary business function. AI users such as company clients and customers are not included in the legal system and framework in which juristic persons are regulated and managed.

If AI should be deemed to be a juristic person, the question arises as to whether the South African Companies Act 71 of 2008 could serve as a guiding legal framework for the drafting of a statute specific to AI regulation. Considering the possibility of this by working through the Companies Act and its 225 sections to see which sections could and could not be effectively applied to AI systems by effecting word amendments is a tedious process, which in depth consideration, arguments and counter-arguments are beyond the scope of this research however serves to highlight potential future research that could and should be undertaken. Although comprehensive consideration and argument of this is beyond the scope of this research, for purpose of demonstrating this practicality a few specific sections of the South Africa Companies Act will be referred too.

It should be noted that although not all of the 225 sections of the Companies Act 71 of 2008 can be applied directly as is to AI systems and its future regulation, the Companies Act does provide a very good basis from which certain sections may be taken as is or which may, subject to slight amendment, such as changing the word “company” ; “companies” or “enterprises” to “AI” or “AI systems” or changing “director” or “directors” to “programmers”, it may nevertheless effectively apply to AI systems and its development, deployment and use. Thereby providing a good basis from which to effectively regulate it in South Africa. By way of example a few of these sections are discussed below:

The preamble of the Companies Act confirms that certain aspects and aims of the Companies Act could also be applied to AI. Relevant parts of the Companies Act preamble confirms that the Act is promulgated for the purpose of providing for the incorporation, registration, organisation and management of companies. Additionally, it serves to define the relationships between companies and their respective shareholders or members and directors. Further, the Companies Act preamble confirms that it provides for efficient rescue of financially distressed companies and provides for appropriate legal redress for investors and third parties with respect to companies. The Companies Act preamble moreover confirms that it establishes a Companies and Intellectual Property Commission and a Takeover Regulation Panel to administer the requirements of the Act itself with respect to companies and to establish a Companies Tribunal to facilitate alternative dispute resolution and to review decisions of the Commission. Lastly the preamble of the Companies Act confirms that it establishes a Financial Reporting Standards Council to advise on requirements for financial record-keeping and reporting by companies.

By replacing the word “company” with “AI” it is evident that the Companies Act preamble could serve to provide the basis of a preamble for a similar AI Act that is specifically promulgated for purposes of: firstly, providing for the incorporation, registration, organisation and management of AI systems. Secondly, it may serve to define the relationships between AI systems and their respective programmers and developers. Thirdly, the Act could provide for efficient rescue of financially distressed AI systems

and provide for appropriate legal redress for investors and third parties (users) with respect to AI systems. Fourthly, the Act could establish an AI systems Commission and Regulation Panel to administer the requirements of the Act itself with respect to AI systems and establish a Tribunal to facilitate alternative dispute resolution and to review decisions of the Commission. Lastly, the Act could establish a Financial Reporting Standards Council to advise on requirements for financial record-keeping and reporting by and on AI systems.

Likewise, Section 2 of the Companies Act which provides for related and inter-related persons and control of companies could be used as a means of determining the relationship between the AI system and its owners, developers and programmers. As provided above, should juristic personality be bestowed on AI systems, the AI system as a juristic person would be owned by another juristic person, the AI holding company, being the company through which owners, programmers and developers operate. Consequently, Section 3 of the Companies Act setting out provisions to regulate and govern subsidiary relations could equally be applied with appropriate amendments to the relationship between an AI system and its relevant owners, developers and programmers.

Similarly, Section 7 of the Companies Act that confirms the purpose of the Act itself could with appropriate amendment of the words “company” and “companies” serve as is, as the purpose for which an AI statute may be promulgated. Taking the wording of section 7 of the Companies Act as is and merely replacing the word ‘company’; ‘companies’ or ‘enterprises’ with the word “AI systems”, the following result is achieved: The purposes of the Act , would like the Companies Act, serve to promote compliance with the Bill of Rights as provided for in the Constitution, in the application of law; promote the development of the South African economy by encouraging entrepreneurship and enterprise efficiency; creating flexibility and simplicity in the formation and maintenance of AI systems; and encouraging transparency and high standards of corporate governance as appropriate, given the significant role of AI systems within the social and economic life of the nation; promote innovation and investment in the South African markets; affirm the concept of the AI system as a means of achieving economic and social benefits; continue to provide for the creation and use of AI systems, in a manner that enhances the economic welfare of South Africa as a partner within the global economy; promote the development of AI systems within all sectors of the economy, and encourage active participation in economic organisation, management and productivity; create optimum conditions for the aggregation of capital for productive purposes, and for the investment of that capital in AI systems and the spreading of economic risk; provide for the formation, operation and accountability of AI systems in a manner designed to promote, support and enhance the capacity of such AI systems to perform their functions; balance the rights and obligations of owners, developers and programmers associated with the AI system; encourage the efficient and responsible management of AI systems; provide for the efficient rescue and recovery of financially distressed AI systems, in a manner that balances the rights and interests of all relevant stakeholders; and provide a predictable and effective environment for the efficient regulation of AI systems [39].

By drawing on the established regulatory framework of the Companies Act, legislators could create a coherent structure for recognising and categorising AI systems as juristic persons as provided in section 8 of the Companies Act. With minor amendments substituting the word “company” with “AI system,” one might conceptualise: for-profit AI systems, that is AI created primarily to generate profit for its developers or programmers, but with limitations on who may participate in ownership. Public AI systems, that are for-profit AI open to participation by the broader public or public institutions, without restrictions on ownership. State-owned AI systems which are AI developed, controlled, and owned by the state, for public benefit and lastly non-profit AI systems, which could be AI created not for profit but to advance community or social interests, potentially qualifying for tax benefits and state subsidies.

Section 13 providing for the right to incorporate company or transfer registration of foreign company can likewise subject to amending the word “companies” to “AI systems” or “Company” to “AI system” be used as the basis on which the right to create an AI system is governed and regulated in a separate AI statute. As required by Section 13 of the Companies Act, a closely worded copy of a similar provision in an AI statute could provide that an AI system may be developed and deployed by any person subject to completion and signature of a prescribed founding documents, such as is served by a company Memorandum of Incorporation. Additionally, such founding document could then be required to be registered with a commission which like the Companies and Intellectual Property Commission are incorporated for purpose of monitoring compliance and enforcement of the AI statute.

Similarly, as section 15 of the Companies Act confirms what provisions may be included in a company memorandum of incorporation, a similar provision could be used to control what provisions may be encoded into the AI system itself confirming what actions by the AI would be permissible and what would be prohibited.

Equally, section 214 of the Companies Act which provides for offence and penalty for false statements, reckless conduct and non-compliance could with appropriate word amendments of “company” to “AI system” serve as the basis on which an AI system deemed a juristic person like a company may be held liable for an offence and breach of the relevant statute. An individual found guilty of an offence in the Companies Act is in terms of section 216 of the Companies Act liable for a fine or imprisonment for 10 years or 12 months depending on the nature of the offence and sections of the Act that has been breached. The same provision may be applied to AI systems and where an individual of the juristic person owning the AI system, that is the juristic person, should be found to have committed an offence the piercing of the corporate veil would be applied to the owning company as juristic person.

Provisions of Companies Act detailing the board of directors and the role and duties of the board of directors cannot find simple application to AI systems as AI systems do not themselves have directors like a company has directors and through which the

company acts. However, this short coming can be rectified as discussed above by assigning the role of shareholders and directors as traditionally understood in terms of corporate law to developers and programmers. By amending the word shareholders as understood in corporate law to developers and amending directors or board of directors in Companies Act to programmers, the relevant provisions of section 66 of the Companies Act that regulates director rights and duties can appropriately govern and regulate the rights and duties of programmers in so far as juristic AI systems are concerned.

Here it is important that a clear distinction is made between developers and programmers. Developers refer to individuals that conceptualise the AI system, referring to individuals that design, architect and oversee the creation of the AI system while programmers are the individuals who write the actual code of algorithms on which the AI system runs. The programmers implement algorithms, structures and functions as per developer instructions to allow AI to process data, learn and produce novel outputs. Consequently, the developers can be seen as the shareholders who invest money, time and effort in conceptualising the juristic person, the company or AI, while programmers can be seen as the directors who control and direct the day to day operations of the AI system as directors do in respect of the company, for and in the shareholders best interests.

Given the above discussions and reference to potentially relevant sections of the Companies Act there is a clear possibility that the remaining sections and provisions of the Companies Act may also with appropriate amendments and adjustments adequately provide for the regulation of AI and AI systems. It is not suggested that the Companies Act can be transplanted wholesale onto AI by merely substituting “company” with “AI.” Rather, the Act is presented as a valuable point of departure for developing a tailored AI statute. The Companies Act demonstrates tested regulatory mechanisms that have proven effective in governing juristic persons, such as accountability structures, disclosure requirements, and liability frameworks. These mechanisms, while not directly transferable, can serve as guiding principles in shaping an AI-specific statute that responds to the unique challenges of AI while drawing on established governance models.

10 Conclusion

As AI in increasingly incorporated into social and corporate life there is an urgent need to consider how it may be appropriately regulated, specifically in how liability may be imposed and compensation claimed where harm, loss or damage should be caused by or through it. A possible solution as proposed in this research essay is to bestow juristic personality on AI in the same manner in which it has been bestowed on companies and similar corporations.

In terms of South African common law requirements, juristic personality can be bestowed on non-human entities that can continue to exist irrespective of changes in its

owners or members, possessing perpetual succession, where it can or is the bearer of rights and duties and when it strives towards a legal, predetermined goal. Given that AI continues to exist as an entity, irrespective of changes in its ownership, programmers and/or developers, has been granted intellectual property rights as has evidenced by the patent awarded to DABUS, confirming the ability of AI to possess property and accompanying rights and duties and strives toward increasing the productivity and efficiency of its users, a legal predetermined goal, it can be concluded that AI may in terms of South African law be bestowed with juristic personality.

By bestowing juristic personality on AI, AI will be deemed to be a hypothetical person separate and independent from its owners, developers, programmers and users. As a juristic person it will have its own rights and duties, own its own assets and be liable for its own debts and liabilities. Liability is enhanced and not reduced by granting juristic personality to AI. Claims for harm, loss or damage can be claimed from the AI itself as a person as opposed to from its owners, developers and programmers, who may evade liability given the collaborative efforts of designing AI and difficulty in tracing back harmful AI actions to one of the many stakeholders involved in designing it. Furthermore, compensation can be paid by the AI directly from the proprietary assets and interests it possesses given its ownership of intellectual property and the royalties associated with same. Additionally existing legal rules, principles and concepts across the various branches of law that already regulate juristic persons can be applied to the regulation of AI. In turn should AI be lawfully regulated it will foster innovation and economic growth, providing for new economic activities and increase employment rates. Given that legal personality has been awarded to non-human entities such as companies, lakes, rivers, temples and eco-systems, there is little room to argue that it cannot be bestowed on AI systems.

Clarity on the legal status of AI is not required but demanded as appropriate regulation of AI will only be possible once there is certainty on AI legal status which will assist in determination of whether existing laws may merely be amended and extended to AI or whether the promulgation of new laws will be required to address the current lack of appropriate AI regulation in South Africa.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Stanford HAI: 2025 AI Index Report (2025). <https://hai.stanford.edu/ai-index/2025-ai-index-report>. last accessed 2025/07/09.
2. Oyekunle, D. & Boohene, D.: Digital Transformation Potential: The Role of Artificial Intelligence in Business. International Journal of Professional Business Review (2024) 9(3).

3. Singla, A., Sukharevsky, A., Yee, L., Chiu, M., Hall, B.: The state of AI: How organisations are rewiring to capture value. Quantum Black AI by McKinsey (2025). <https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai/#.> last accessed 2025/07/23.
4. Annon: SA Organisations Embrace AI. (2024) World Wide Worx available at <https://www.worldwideworx.com/sagenairoadmap/> accessed 3 October 2025
5. Communications & Digital Technologies ‘South African National Artificial Intelligence Policy Framework’ Department: Communication & Digital Technologies Republic of South Africa available at <https://www.dcdt.gov.za/sa-national-ai-policy-framework/file/338-sa-national-ai-policy-framework.html>, accessed 28 August 2025.
6. Communications & Digital Technologies Department ‘National AI Summit’ Government Department: Communications & Digital Technologies Republic of South Africa (2023) available at <https://www.dcdt.gov.za/topics/495-national-ai-summit.html>, accessed 28 August 2025.
7. Annon: AI Watch:Global Regulatory Tracker- South Africa. White & Case available at <https://www.whitecase.com/insight-our-thinking/ai-watch-global-regulatory-tracker-south-africa> accessed 3 October 2025.
8. Parker v Forsyth NNO and Others (1585/20) [2023] ZAGPRD 1 (29 June 2023); Mavundla v MEC: Department of Co-Operative Government and Traditional Affairs KwaZulu-Natal and Others 2025 (3) SA 534 ; Northbound Processing (Pty) Ltd v South African Diamond and Precious Metals Regulator and Others (2025/072038) [2025] ZAGPJHC 661 (30 June 2025)
9. Marshall, B.: Opinion: No Legal Personhood for AI. Patterns (2023), <https://www.sciencedirect.com/science/article/pii/S2666389923002453>, last accessed 2025/07/24.
10. Bharti, A., Kaur, G.: Exploring the Legal Personality of Artificial Intelligence: Challenges, Opportunities, and Future Directions. Journal of Informatics Education and Research 5(1), 2483 (2025).
11. Lehrberg, B.: AI as Juristic Person. De Lege (2021), 52. <https://journals.uu.se/DeLege/article/view/416/371>, last accessed 2025/07/23.
12. Neumann, O., Guirguis, K., Steiner, R.: Exploring artificial intelligence adoption in public organizations: a comparative case study. Public Management Review (2022), 3. <https://doi.org/10.1080/14719037.2022.2048685>, last accessed 2023/06/07.
13. Brown, R.D.: Property ownership and the legal personhood of artificial intelligence. Information & Communications Technology Law 30(2), 211 (2021); Mertens, F.: Financial Law Institute: Faculty of Law and Criminology Working Paper Series 5 (2023).
14. Jose, N.S.: Navigating the Legal Landscape: Evaluating the Case for Artificial Intelligence as Juristic Persons. Indian Journal of Computer Science (2023), 33–35.

15. Tiwari, T; Tiwari, T & Tiwari, S : How Artificial Intelligence, Machine Learning and Deep Learning are Radically Different?' International Journal of Advanced Research in Computer Science and Software Engineering (2018) 8(2);
16. Annon 'What is AI' MetaLab at Harvard available at <https://aipedagogy.org/guide/resources/> accessed 27 August 2025.
17. Dr. Shrivastava, A. ; Pandey, A. ; Singh, N.; Srivastava, S.; Srivastava, M.; Srivastava, A.: Artificial intelligence(AI): Evolution, Methodologies, and Applications' International Journal for Research in Applied Science & Engineering Technology (2024) 12(4)
18. Anon.: What is Machine Learning. IBM, <https://www.ibm.com/zazen/cloud/learn/machine-learning>, last accessed 2023/02/21.
19. Surden, H.:ChatGPT, Large Language Models, and Law' Fordham L. Rev (2024) 92(5) at 1941 - 1972
20. Håkansson, A. & Philips-Wren, G.: Generative AI and Large Language Models- Benefits, Drawbacks, Future and Recommendations' Procedia Computer Science (2024) 246;
21. Annon 'What are large language models (LLMs)' IBM available at <https://www.ibm.com/think/topics/large-language-models> accessed 29 August 2025.
22. Blank, I.A.:What are Large Language Models Supposed to Model' Trends in Cognitive Sciences (2023) 27(11);
23. Stöffelbauer, A.: How Large Language Models Work Medium (2023) available at <https://medium.com/data-science-at-microsoft/how-large-language-models-work-91c362f5b78f> accessed 29 August 2025.
24. Linkon, A.A; Shaima, M.; Sarker, S.U.; Badruddowza ; Nabi, N.; Rana, N.U.; Ghosh, S.K.; Rahman, M.A.; Esa, H. & Chowdhury, F.R.: Advancements and Applications of Generative Artificial Intelligence and Large Language Models on Business Management: A Comprehensive Review' Journal of computer Science and Technology Studies (2024)
25. Annon 'Introduction to large language models' Google available at <https://developers.google.com/machine-learning/resources/intro-llms> accessed 29 August 2025.
26. Stryker, C. & Kavlakoglu, E.: What is AI IBM available at <https://www.ibm.com/think/topics/artificial-intelligence>, accessed 26 August 2025.
27. Forrest, K.B : The Ethics and Challenges of Legal Personhood for AI. The Yale Law Journal Forum (2024)
28. Anon " What is Machine Learning" <https://www.ibm.com/zazen/cloud/learn/machine-learning> (accessed 21-02-2023)
29. Stryker, C. & Holdsworth, J: What is Natural Language Processing. IBM, <https://www.ibm.com/topics/natural-language-processing#:~:text=the%20next%20step-,What%20is%20natural%20language%20processing%3F,same%20way%20human%20beings%20can>, last accessed 2023/02/20.

30. Kerikmae, T.; Muursepp, P.; Pihl, H.M.; Hamulak, O. & Kocharyan, H.: Legal Person – or – Agenthood of Artificial Intelligence Technologies Acta Baltica Historiae et Philosophiae Scientiarum (2020) 8(2)
31. Yalman, B.: Electronic Personhood: A Compact Analysis of Legal Personality for Artificial Intelligence. *ex/ante* (2024), 3–13. https://doi.org/10.3256/978-3-03929-059-8_02.Author, F.: Article title. Journal 2(5), 99–110 (2016)
32. Novelli, C.; Floridi, L.; Sartor, G. & Teubner, G.: AI as Legal Persons – Past, Patterns, and Prospects (2024). forthcoming in the Journal of Law and Society, Available at SSRN: <https://ssrn.com/abstract=5032265> or <http://dx.doi.org/10.2139/ssrn.5032265>
33. Lovell, J.: Legal Aspects of Artificial Intelligence Personhood: Exploring the Possibility of Granting Legal Personhood to Advanced AI Systems and the Implications for Liability, Rights and Responsibilities (2023), SSRN, <https://ssrn.com/abstract=4749785>, 5.
34. Chesterman, S.: Artificial Intelligence and the Limits of Legal Personality. International and Comparative Law Quarterly 69, 819–844 (2020), at 822.
35. Papakonstantinou, V., de Hert, P.: Refusing to award legal personality to AI: Why the European Parliament Got it Wrong. European Law Blog/AI Regulation (2020), <https://www.europeanlawblog.eu/pub/refusing-to-award-legal-personality-to-ai-why-the-european-parliament-got-it-wrong/release/1>, last accessed 2025/07/25.
36. Davel, C.J., Jordaan, R.A.: Law of Persons. Juta & Co Ltd, South Africa (2025), 4.
37. Oxford English Dictionary. https://www.google.com/search?q=entity&rlz=1C1VDKB_enZA1026ZA1026&oq=entity+&gs_lcp=EgZjaHJvbWUyBggAEEUYOTIHCAEQABiPAtIBCDExMjdqMGo5qAIA-sAIB&sourceid=chrome&ie=UTF-8 accessed 31/07/25
38. Thaldar, D., Naidoo, M.: AI inventorship: The right decision? South African Journal of Science 117(11–12), 1–3 (2021). <https://doi.org/10.17159/sajs.2021/12509>
39. Section 7 Companies Act 71 of 2008 – amended to include word AI

Digital Colonialism as an Economic Strategy: Engineered Inequality

Sunita Menon¹²

¹ University of Johannesburg, 5 Kingsway Ave, Rossmore, Johannesburg, 2092

² United Nations University, 5-chōme-53-70 Jingūmae, Shibuya, Tokyo 150-8925, Japan
sunitam@uj.ac.za

Abstract. While the systemic biases and discriminatory outcomes of AI are increasingly well-documented, this article argues that these dynamics also facilitate a subtler, less examined form of neo-colonialism that disproportionately affects Africa due to enduring historical and structural power imbalances in the global digital ecosystem. A pertinent question, however, is whether unethical biases are deliberate or unintended. This paper examines the intrinsic algorithmic biases that disproportionately marginalize communities in Africa, critically analyzing whether these computational systems replicate and reinforce entrenched historical power asymmetries, and to what extent this occurs intentionally, through the lens of postcolonial theory. Through an analysis of ownership structures, capital flows, and policy decisions in AI development and deployment, this paper demonstrates that digital discrimination reflects both structural reproduction of inequality and increasingly deliberate institutional choices. While some bias emerges from inherited systemic patterns, the systematic elimination of equity safeguards and declining of inclusive AI commitments by actors with full knowledge of consequences reveals that exclusion has become a foreseeable outcome of deliberate policy decisions within technological ecosystems that reflect and reproduce colonial-era hierarchies of power and knowledge.

Keywords: AI, AI in Africa, algorithmic bias, digital discrimination, colonialism

1 Introduction

The South Korean economist Ha-Joon Chang asserts: “Economics has become like Catholic theology in medieval Europe. It has become the language of the rulers... they are not going to let you speak it unless you belong to the priesthood or are from an extremely privileged background.” [1] In essence, Chang’s statement highlights how the exclusivity of economics can perpetuate systemic biases and discrimination, limiting the field’s ability to serve the broader public interest. The insularity and complexity of contemporary economics serve as barriers to democratic deliberation and social

change just as Latin and theology became exclusionary. If artificial intelligence (AI) is central to the current economic system, then the biases embedded in AI algorithms will perpetuate exclusion. This paper demonstrates the exclusionary nature of AI systems by providing a case study on racial bias, revealing how algorithmic decisions can systematically disadvantage certain groups. Racial bias refers to the systematic and often unintended discrimination that occurs when AI systems produce outcomes that disadvantage individuals or groups [2]. Based on existing capital and ownership structures, an argument is then made that this bias may not be incidental but rather a deliberate tool to perpetuate new and pervasive forms of colonialism.

2 AI, the Global Economy and Neo-Colonial Models

AI has emerged as a transformative force in the global economy. Its applications have revolutionized entire industries and offered unprecedented opportunities for innovation and human advancement. There are compelling arguments that it has now become the central driver of the global economy. AI is projected to contribute \$15.7 trillion to global GDP by 2030 [3]. However, with the current status quo, this projection is expected to have a significantly smaller impact on Africa. Collectively Africa, Oceania, and some less developed Asian markets could grow by just \$1.2 trillion or 5.6% GDP growth [4]. Marwala [5] argues that Africa is increasingly viewed as a passive actor because it is not innovating on the scale of the US or China, for example. This assertion provides a context for understanding why the continent has been forced into a dependency role which is counter to the developmental needs of the continent.

This disparity is not only economic but structural. Inequity in this regard is not just confined to delayed adoption or a lack of infrastructure but also reflects deeper structural inequalities that shape the development, deployment, governance and even impact of these technologies [6] [7]. AI carries deeply problematic undercurrents of systemic bias that lead to digital discrimination resulting in a new form of colonialism that disproportionately and negatively impacts Africa because of a persistence of historical power imbalances. While bias, in its simplest sense, denotes a tendency or inclination within data that can produce either favorable or unfavorable outcomes depending on its context, this paper specifically examines biases that result in unfair or unethical consequences. Although the primary focus of this paper is racial bias in particular, it acknowledges that bias can intersect with other factors, such as gender and class and thus cannot be viewed in a silo. Menon [8] asserts that the assumed neutrality of AI systems is a misrepresentation as there are both overt and covert biases embedded in the algorithms that inform these systems. In terms of racial bias, she argues that Africa and the African diaspora are disproportionately impacted by this misrepresentation. These biases thus create new and pervasive colonial pathways resulting in the “algorithmic colonization of Africa” [9]. Such harms are not abstract and range from biased facial recognition systems to data extraction practices that replicate exploitative economic models [10]. The import of this argument is that power and influence exerted by the West is

reminiscent of historical patterns of colonialism. The implications for Africa cannot be underplayed as it strips citizens of autonomy, can be exploitative and could erode local cultures and identities [9]. There is an argument to be made that this new form of colonization has led to developmental challenges in Africa, which if unchallenged will persist [9]. Within digital spaces, challenges to sovereignty, economic exploitation and cultural imposition take place [8]. Given the current power dynamics and concentration of AI, Priya and Arockiasamy [11] caution that developing countries may become dependent on AI technologies developed elsewhere thus perpetuating a cycle of technological and economic subordination. This could be further exacerbated by the prioritization of Western values and norms in these systems. The emerging challenges of algorithmic bias could undermine the very principles of autonomy and justice. This debate draws parallels between historical colonization practices and the current digital landscape which is controlled and powered by the West.

While the rhetoric around bias is usually centered around divisions between the Global North and the Global South, there is merit in honing into Africa specifically. While Kloß [12] and Levander and Mignolo [13] argue that the Global South is a political and ideological construct that emerges in response to global inequalities and power struggles, there are significant disparities within this grouping. Prashad [14] and Kloß [12] thus caution against viewing the Global South as a homogeneous entity. The degrees of separation are apparent in understandings of AI power dynamics. AI power concentration lies in the West and in China to a slightly lesser extent. Recent estimations, however, indicate that China is quickly closing the gap, particularly following the launch of DeepSeek's generative model in January 2025. Lee [15] states: "Previously I think it was a six to nine month gap and behind in everything. And now I think that's probably three months behind in some of the core technologies, but actually ahead in some specific areas." This indicates that the Global South cannot be viewed as a single unit as there are differing levels of AI development and deployment with areas in Asia emerging as significant global competitors. Africa thus emerges as a unique case study in any exploration about algorithmic bias. Africa has been referred to as the center of the Global South but as Yeats writes in 'The Second Coming', "Things fall apart; the center cannot hold." [16] While Africa is positioned as the center, there are forces both external and internal that threaten to make real the fate Yeats describes as "mere anarchy is loosed upon the world" This has proved an accurate sentiment in the case of Africa's AI development and deployment. Tran [17] argues: "While there are questions about whether some countries—such as China or Russia—should be considered parts of the Global South (GS), it is obvious that Africa is at the center of the group. Different aspects of Africa—its potential, its reality, and its efforts to realize its potential—embody the challenges and the prospects of the GS in general. More specifically, the difficulties Africa faces, how it will deal with them, its progress or lack of progress, and the changes it would like to see in the current international economic and financial system to help it overcome the obstacles to development, help make clear what the GS is all about." The position adopted is that Africa is central to the Global South, encapsulating both the challenges and prospects faced as its experiences reflect broader issues of development and economic growth. The continent's efforts to overcome its difficulties and navigate the international economic system provide insight

into the core concerns and aspirations of the Global South as a whole. An example of this is Africa's marginalization during the global COVID-19 vaccine rollout. While wealthier nations in the Global North hoarded vaccine doses, African countries were left to rely on donations or delayed shipments. This experience prompted calls for greater health sovereignty as African nations called for patent waivers. In this way, African nations were able to articulate Global South concerns around equity and access [18].

Africa's vulnerability globally and within the Global South thus raises an important question with regards to whether these biases within algorithms are deliberate or unintended. There is an argument to be made that existing uneven power structures enable bias and thus discrimination. While overt malice may not always be present, the persistence of unequal global power structures suggests that bias is not simply accidental, it is enabled and perpetuated by those who hold power. Control of AI is confined to a few large tech companies prompting the question, who controls the development and deployment of AI and consequently, who benefits from its value [19]? Menon [8] calls for the critical examination of AI's technological and power structures to prevent perpetuating historical inequalities. Although somewhat dated considering the rapid proliferation of AI and the extraordinary jump in adoption during the COVID-19 period, the Center for Security and Emerging Technology (CSET) provides valuable information on capital concentration and there seems to be little change in this regard [20]. As of 2019, the United States led the global AI investment market, attracting \$25.2 billion (64% of global AI investment) across 1,412 transactions, with total global AI equity investment reaching \$40 billion across 3,100 transactions. While China's AI investment surged between 2015 and 2017, it declined to near-2015 levels by 2019 attributed primarily to overexpansion in previous years [21]. In the same period, AI investment in Western Europe, Israel, India, Japan, and Singapore grew rapidly. More recently, a 2024 Ascendix report found that the US dominates AI startup funding with an estimated 5,509 AI startups, which attracted \$47 billion in non-governmental funding in 2022. China follows with 1,446 AI startups and \$95 billion in private AI investments that year. China is certainly seeing a rebound since its 2019 slump. As Morgan Stanley [22] phrased it, "a sleeping giant awakens". This is driven by strong government support and a focus on efficiency and cost-effectiveness. Despite challenges from US restrictions, China is pursuing AI self-sufficiency within five years and adopting open-source models. The UK ranks third, with its AI market projected to reach \$1 trillion by 2035. OpenAI leads global AI funding with \$11.3 billion, followed by Anthropic at \$7.7B and Databricks at \$4B. These firms receive significant backing from tech giants such as Microsoft, Google, and Amazon. Corporate capital spending on AI infrastructure is immense, with Google, Microsoft, Meta, and Amazon investing \$52.9 billion in the second quarter of 2024. This has increased significantly in 2025. The combined profits of these companies surged to \$92.17 billion in the second quarter of 2025. Infrastructure investments are also projected to total \$350 billion by the end of this year [23]. With the likes of DeepSeek and humanoid robots emerging, China could very well be considered back in the race. It is projected that China's core AI market could reach \$140 billion by 2030. AI could therefore contribute an extra 0.2–0.3 percentage point to annual GDP growth over the next few years.

While these statistics provide insight into the concentration of power globally, there are further insights into the racial makeup of these companies that indicate why bias is ingrained in these systems. Research from 2022 [24] found that only 25% of AI solution developers, on average, come from racial or ethnic minority groups while 29% of organizations have no minority employees working on AI solutions at all. Despite these damning figures, only a third of these companies have targeted programs to increase this number. Even then, these programs are disproportionately targeted at improving just gender diversity while largely overlooking ethnic diversity. This suggests that bias, and racial bias in particular, may not be accidental or incidental. Existing power structures don't just enable bias, but potentially manufacture it systematically. For example, three major facial-analysis programs were found to have significant skin-tone and gender biases, with error rates under 1% for light-skinned men but up to nearly 47% for darker-skinned women [25]. The study by MIT and Stanford researchers [25] demonstrates how training data dominated by white, male faces skews AI accuracy. Their findings raise concerns about systemic discrimination in commercial AI systems. The study concludes that existing power structures not only overlook marginalized groups but embed their exclusion into supposedly objective technologies. This bias thus isn't merely a technical flaw but rather a consequence of systemic decisions about whose data is collected and valued and whose experiences are rendered invisible in technological design. These outcomes reflect global hierarchies of race, gender, geography, and capital. Bias is thus not incidental but institutionalized. As this paper demonstrates, ownership and control become crucial indicators of potential deliberate discrimination. As control of AI development is centralized in Western (primarily US) and Chinese tech ecosystems, these companies have minimal representation from or accountability to African contexts. As the ownership metrics and investment figures show, the US and China constitute the twin poles of AI innovation and infrastructure. These powers hold data, talent, computing power as well as the geopolitical leverage to shape global AI trajectories. Africa could thus be considered enslaved by these AI powers as it sits on the periphery and the receiving end of AI. The outsider phenomenon amplifies Africa's marginalization, where its people are subject to external systems and biases without meaningful participation or agency in the creation or oversight of these transformative technologies. The lack of diverse development teams inherently embeds particular worldviews and assumptions. While there is little information on the exact makeup of these teams, the 2025 turnaround on diversity, equity, and inclusion (DEI), which the White House [26] has termed "shameless discrimination" is certainly an indictment. Furendal [19] argues that the prevailing private ownership model also poses systemic challenges that cannot be effectively addressed through regulation alone, indicating a pressing need for alternative ownership models to ensure more equitable control. This very clarion call stands in opposition to current policies. The scrapping of DEI by the Trump administration does not bode well for shifting dynamics in the US. This could stand in direct opposition to the argument for localized systems considering where capital flows stem from. A Forbes report [27] found that as of 24 March 2025, at least 39 major companies have rolled back or eliminated their DEI initiatives. The technology thus creates value for developers and owners and those that look like them and not necessarily for those experiencing its effects. Calvin and Leung [28] suggest that the

current strategies employed by AI developers can lead to exclusionary practices. For example, while many AI companies are patenting their technologies to protect against infringement lawsuits and to maintain a competitive edge, they also share research to attract talent and customers. This dual approach raises concerns about how open they are willing to be. Furthermore, the presence of ‘patent trolls’ or entities that hold patents solely for litigation purposes without producing any products can create an environment where companies may become more secretive about their innovations to avoid lawsuits. In January 2025, the White House [29] also released an executive order aimed at reinforcing the United States’ leadership in AI by ensuring its development remains free from ideological bias and regulatory barriers. Then, in February, both the US and the UK declined signing a ‘Statement on Inclusive and Sustainable AI’ at the Paris AI Action Summit. At the summit vice president JD Vance [30] declared that the US’s priority is competitiveness, adding: “The AI future is not going to be won by hand-wringing about safety.” Consequently, there are concerns that AI companies supplying the US government may face pressure to remove safety measures if they are seen as supporting DEI or hindering innovation, which could lead to fewer safeguards in AI products. Bishop [31] argues: “All signs suggest the Trump administration favors a reduction in the ethical regulation of AI. The executive orders may be interpreted as allowing or encouraging the free expression and generation of even discriminatory and harmful views on subjects such as women, race, LGBTQIA+ individuals and immigrants.”

These examples reflect broader concerns about the alignment of unaccountable technological power with state interests. The growing influence of tech executives in shaping both public policy and military applications represents surveillance capitalism’s capture of democratic institutions [32]. In this scenario, private economic interests increasingly determine public technological trajectories. As governments eliminate safety frameworks while simultaneously expanding military AI investments, it raises concerns about authoritarian applications of AI [33]. These applications erode democratic norms as power is centralized through opaque corporate-state partnerships that operate beyond meaningful public oversight [34].

A fitting example of the potential dangers of this approach is the chatbot Tay that was launched by Microsoft in 2016 to learn human speech patterns through Twitter interactions [35]. However, it was shut down within 24 hours after generating racist, sexist, and anti-Semitic tweets. In response, Microsoft [36] released a statement stating: “Looking ahead, we face some difficult – and yet exciting – research challenges in AI design. AI systems feed off of both positive and negative interactions with people. In that sense, the challenges are just as much social as they are technical. We will do everything possible to limit technical exploits but also know we cannot fully predict all possible human interactive misuses without learning from mistakes.” Zemčík [37] asks if the failure of Tay was inherently malicious, flawed, and ineffective, or whether our judgment is influenced by cognitive shortcuts and biases. The author argued that the chatbot “held up a mirror to people about how things were in reality—as if it had a mind of its own.” Neff and Nagy [38] assert that Tay’s responses reflected “social relationships and the state of humanity”. This example seems to have been less of a warning than a case study in unregulated and unchecked AI. This year, Elon Musk’s AI chatbot, Grok 3, has made unfiltered, controversial remarks on politics, free speech,

and social issues, particularly in India. In response, it declared [39]: “some slam me for bias, others cheer.” Grok also reinforced the white genocide narrative in South Africa by perpetuating far right and white propaganda that had no empirical basis. The generative model has gone so far as to term itself ‘Mecha-Hitler’. In the aftermath, its developer, xAI, secured a \$200 million contract with the U.S. Department of Defense. Zemčík [37] concludes that AI can both mitigate and introduce cognitive biases, but it is important to note that these systems lack morality, emotions, and consequences for their actions even as users attribute human-like accountability to them. Vorsino [40] postulates that Tay is located in web 2.0, which is influenced by neoliberal, racialized, and gendered frameworks that define its structure.

There is thus justified concern about how the economic power amassed by large AI companies can translate into political influence, which could undermine democratic processes as seen in the example of Grok 3’s political leanings [19]. This is demonstrated by the potential for AI companies to lobby lawmakers, influence legislation in their favor, and limit the capability of citizens to effect collective decisions [19]. Bias could thus be considered a feature, not a bug, of current technological ecosystems. Increasingly so, it demonstrates how colonial-era power dynamics are technologically reproduced. Moreover, as these ownership structures indicate, there is also disparity within the Global South. Ballim and Breckenridge [41] ask a crucial question: “African people, firms and societies have produced, have been monopolized and discounted by metropolitan corporations with the energetic assistance of local elites. Will the growing power of the centers of AI in the United States and China – and the global monopoly power of a small number of firms secured by AI – produce a new era of data-driven extraversion and dependency?”

There is also an acknowledgement from the US that China is a formidable player in AI innovation and thus a significant strategic competitor, according to the Final Report of the National Security Commission on Artificial Intelligence [42]. Intriguingly, China, the other “AI superpower” was a signatory to the ‘Statement on Inclusive and Sustainable AI’. However, China’s largely inward AI policies, which are focused primarily on domestic development and national security may hinder hopes for inclusivity. Although written in 2018, Kai Fu-Lee’s [43] AI Superpowers: China, Silicon Valley, and the New World Order makes the argument: “The West may have sparked the fire of deep learning, but China will be the biggest beneficiary of the heat the AI fire is generating.” The US report makes an argument for the US to partner with historical allies of the United States, such as nations in Europe (e.g., the United Kingdom, Germany, and France) and partners in Asia (such as Japan, South Korea, and Australia) to work together on AI standards, research, and security. The report also highlights the potential for cooperation with emerging AI powers that share democratic values, including countries such as India and Canada as these nations are seen as vital partners in the global AI landscape. There is unsurprisingly no mention of any African nations, indicating that power is indeed located elsewhere. Moreover, China’s growing technological presence in Africa has also sparked concerns about digital neo-colonialism, which allows for the exertion of control and influence over African nations [44]. A potential future could thus be a Chinese-led AI landscape.

Although China has a rather comprehensive set of AI regulations that advocate for inclusivity and transparency, there are also accompanying criticisms. Discussions on Chinese AI governance often overlook the substance and policymaking process of its regulations. Commentary typically falls into two extremes: dismissing the regulations as insignificant or using them for political leverage. Critics argue that since Xi Jinping and the CCP can override their own rules, these regulations hold little real importance [45]. In a global context, despite technological prowess and advancement, this once again indicates the glaring divide between the Global North and the Global South. This concentration of AI power, with Africa systematically excluded from its development and governance, raises fundamental questions of technological sovereignty and postcolonial power relations

3 A Postcolonial Framework

Menon [8] and Marwala [5] make the argument for inclusive and localized datasets to reimagine the resulting power imbalance. However, the ownership structures and capital flows in AI development and deployment stand in direct opposition to this goal. Despite having over 2,400 AI companies, Africa attracted only \$4 million in AI funding across five deals in Q2 2024, compared to \$23.2 billion raised globally. This means less than 1% of global AI investment flows into Africa [46].

Western-developed AI is fundamentally misaligned with African contexts and interests thus reproducing power imbalances by maintaining technological dependency. Africa has been portrayed as a ‘passive observer’ in technological development, which is an ideologically flawed narrative. Colonial interruption has systematically excluded African scholars, artisans, and scientists from developing indigenous technological solutions. The perceived technological lag is actually a result of colonial disruption, not inherent African inability. Postcolonial theory provides a crucial framework for understanding these power differentials as the concept of the ‘other’ helps interpret how algorithmic bias continues colonial-era discrimination [8]. The technology represents a new form of discrimination that operates subtly across all societal spheres. Without addressing these systemic biases, Africa’s socio-economic development will remain constrained. Nkrumah [47] cautioned that neocolonialism would represent the ultimate phase of imperialism and that it would hinder progress and reverse development. The policies put in place surrounding AI and the emerging inequities in the Global South make a compelling argument for a neocolonial bend. Priya and Arockiasamy [11] state that a: “sort of humiliation and underestimation on Blacks is continued in the Postcolonial condition too. The Western domination perpetuates through AI tools.”

There is an argument to be made for decolonial justice-centered alternatives. In this regard, Menon [8] highlights the need for interventions such as policy development, ethical charters, diverse datasets, and regulatory oversight bodies to address digital discrimination through a postcolonial lens. The shift, however, globally in terms of AI governance etc. remains in question and challenges the very notion of an effective post-colonial strategy. Calvin and Leung [29] suggest AI companies adopt patent pools where companies might share their patents to reduce litigation risks and promote inno-

vation. If structured correctly, this could mitigate some exclusionary practices by allowing broader access to technology.

Munyua [48] demonstrates that successful multistakeholder engagement in internet governance is possible in African contexts by documenting Kenya's inclusive ICT policy development process from 2003-2006. However, this success has not translated to broader continental coordination. The 2021 crisis at the African Network Information Center (AFRINIC), the regional internet registry responsible for IP address allocation across Africa, revealed systemic governance failures. Despite AFRINIC's mandate to promote meaningful state participation in internet governance, African governments remained largely absent from its governance structures, even though they had access to mechanisms like the Africa Government Working Group (AfGWG) and membership in ICANN's Governmental Advisory Committee (GAC) [49].

This disengagement from regional internet infrastructure governance reflects deeper institutional failures. In the multistakeholder model of global internet governance, governments play a vital role in upholding the public interest, particularly in areas such as cybersecurity, development, and cross-border data regulation [50]. In Africa, this role has often been either neglected or distorted by attempts at centralized control [49]. While Western dominance of global internet governance structures creates structural imbalances, the failure of African governments to engage consistently has compounded these challenges. This dual dynamic of external power asymmetries combined with internal governmental inertia enables Western influence to persist largely unchallenged [51].

Effective participation in internet governance requires structure, continuity, capacity and sustained political will [52]. Without these elements, governance vacuums are filled by either private interests or external actors. Moving forward requires structured engagement with clear boundaries, capacity building to ensure policymakers across sectors understand internet governance issues, continental coordination mechanisms, and a focus on public interest as the guiding principle for government involvement [50].

4 From Historical to Digital Colonialism: A Conceptual Framework

To understand whether AI bias is deliberate requires an understanding of colonialism and how it continues to operate today. Colonialism is not merely a historical period that ended with formal independence but an ongoing structural system that still shapes global power relations [53]. The colonial process, concentrated primarily in the 19th and 20th centuries, involved Western powers dominating and exploiting territories while imposing Western constructs and standards on non-Western societies. Yet, the deeper legacy lies in the coloniality of power or the enduring hierarchies that persist in how knowledge is valued, whose labor is exploited and where wealth accumulates [53]. Colonialism restructured economies to serve colonial interests by ensuring that wealth and resources flowed outward while colonized populations were locked into positions of dependency [54]

[55]. In this process, it devalued indigenous knowledge systems by positioning Western epistemologies as universal and objective while dismissing other ways of knowing as backward or irrelevant [56] [57]. These patterns didn't disappear with independence but rather embedded themselves in economic relationships and knowledge hierarchies that shape the contemporary world.

As this paper demonstrates, contemporary AI development operates within and reinforces these colonial structures. African user data flows to Western servers, generating billions in value captured by foreign corporations, while the continent receives less than 1% of global AI investment despite hosting 2,400 AI companies [46]. This mirrors historical patterns where African resources enriched distant centers of power. Algorithmic systems designed in Silicon Valley and Beijing determine creditworthiness or employment decisions, for instance. It could thus be argued that governance is imposed from the outside where it has been optimized for other contexts and populations. African institutions increasingly depend on technologies they didn't build and cannot modify to include contextually appropriate dimensions [9]. Training datasets systematically exclude or misrepresent African knowledge systems and languages thus positioning Western data and norms as the universal standard against which all others are measured and often found deficient [8]. The coloniality of power persists as Western constructs are imposed as technological standards, wealth flows outward and structural dependency deepens [53].

A critical question emerges: are these patterns inevitable consequences of how technology develops or do they reflect choices made by those with power? Systemic bias operates through normalized practices and inherited inequalities. Galtung [58] argues that this is a form of structural violence where harm occurs without requiring intention. Training data captures patterns of discrimination already present in historical records or hiring decisions, for instance. Development teams lack diversity because educational pipelines reflect longstanding barriers to access in STEM fields. Market concentration perpetuates itself as established players leverage network effects and accumulated resources to maintain dominance. These dynamics explain some bias as structural rather than intentional.

However, deliberate bias emerges when institutional actors make policy decisions with foreseeable discriminatory consequences while possessing both knowledge and power to choose otherwise. Consider that 64% of global AI investment concentrates in the United States [20] within firms where only 25% of developers come from racial minorities and 29% of organizations employ no minority AI developers at all [24]. Initially this appears structural as it could be argued that this is the outcome of historical patterns. Yet, when these same firms systematically eliminate diversity initiatives [26], when governments issue executive orders removing equity protections from AI development and when nations explicitly decline signing commitments to inclusive AI, we witness active choices to maintain exclusionary patterns [26] [39]. The difference between structural and deliberate is dependent on conscious policy decisions that predictably perpetuate exclusion.

Some scholars argue AI's geographic concentration reflects natural market dynamics. Porter [59] and Audretsch and Feldman [60] demonstrate that innovation clusters emerge through knowledge spillovers and agglomeration economies. In other

words, physical proximity enables collaboration and rapid iteration. This suggests concentration may be economically efficient rather than discriminatory. Others contend algorithmic bias represents technical complexity or what Selbst et al. [61] term “abstraction traps”. This refers to well-intentioned systems that produce harmful outcomes through design limitations rather than malice.

While these explanations certainly have merit, they cannot account for the policy environment. Historical power asymmetries determined which regions developed enabling institutions in the first place [62] indicating that ‘natural’ clustering reflects inherited advantage rather than pure merit. More importantly, efficiency arguments fail to explain why 39 major companies eliminated diversity initiatives within months [26] or why governments actively dismantled safety frameworks while expanding military AI investments [29]. When institutional actors with full awareness systematically remove equity safeguards, we move beyond unintended consequences to deliberate choices.

The concentration of AI capital and power thus reflects not merely market forces but political decisions about whose innovation merits funding and whose knowledge systems count as valid. This analysis employs policy analysis and political economy frameworks to trace how ownership structures, capital flows, and policy decisions reveal power holders actively maintaining exclusionary systems while possessing both knowledge of consequences and capacity for different choices.

5 Conclusion and Limitations

It is acknowledged that this context is continually changing. While this critique looks at the period in the first half of 2025, a limitation may very well be the fast-shifting nature of both AI, policy and geopolitics. Many of these changes could not have been anticipated just a year prior. Methodologically, this analysis traces patterns through policy documents, investment data, and institutional decisions. While this approach cannot definitively prove individual motivations, it reveals institutional patterns where actors with knowledge and power make choices with foreseeable discriminatory consequences. Additional limitations include restricted access to internal corporate decision-making processes and the time lag between policy implementation and measurable outcomes.

The ascendancy of AI in the global economy has surfaced new fissures of inequality that astoundingly mirror historical patterns of exclusion and power inequalities crucially along lines of race. This paper argues that the ostensibly neutral architectures of AI often reproduce and encode racialized hierarchies, functioning almost as deliberate tools to deepen existing hierarchies of capital, race and power. The evidence demonstrates how exclusion operates through capital concentration in Western firms lacking diversity, the systematic elimination of equity safeguards via DEI rollbacks and safety framework removals, and the creation of structural dependencies as African institutions rely on AI technologies they neither control nor can modify for local contexts.

The question is whether the persistent marginalization of Africa in AI development and governance is accidental. The evidence suggests it reflects broader patterns of digital colonialism that position the continent as a passive recipient. While some algorithmic bias emerges from structural reproduction of existing inequalities, the pattern of deliberate policy choices by institutional actors with full awareness of consequences reveals intentionality [26] [27] [29] [30].

Unless these exclusionary practices are confronted head on through deliberate and intentional shifts in ownership models, regulatory frameworks and capacity-building initiatives, AI will be fundamentally compromised for the Global South – and Africa in particular. The current trajectory of AI carries the inherent risks of further entrenching a new wave of imperialism. There is thus an urgent need for Africa to assert its agency or fall into the trap of yet again a period of engineered inequality.

References

1. Lari, T.: Economics, Pluralism, and Democracy: An Interview with Ha-Joon Chang. *Erasmus Journal for Philosophy and Economics* 17, 21–38 (2024). <https://doi.org/10.23941/ejpe.v17i2.920>
2. Cuella, N.G.: Unconscious Bias: What is Yours? *Journal of Transcultural Nursing* 28, 427–427 (2017). <https://doi.org/10.1177/1043659617710940>
3. Brookings Institution: The future is intelligent: Harnessing the potential of artificial intelligence in Africa. <https://www.brookings.edu/articles/the-future-is-intelligent-harnessing-the-potential-of-artificial-intelligence-in-africa>. Accessed 7 Aug 2025
4. PwC: The Fearless Future: 2025 Global AI Jobs Barometer. <https://www.pwc.com/gx/en/issues/artificial-intelligence/ai-jobs-barometer.html>. Accessed 7 Aug 2025
5. Marwala, T.: Closing the Gap. 1st edn. Pan Macmillan, Johannesburg (2020).
6. Sampath, P.G.: Governing Artificial Intelligence in an Age of Inequality. *Global Policy* 12, S6, 21–31 (2021). <https://doi.org/10.1111/1758-5899.12997>
7. Himmelreich, J., Lim, D.: Artificial Intelligence and Structural Injustice: Foundations for Equity, Values, and Responsibility. In: Bullock, J.B. (ed.) *The Oxford Handbook of AI Governance*, pp. 210–231. Oxford University Press, Oxford (2022).
8. Menon, S.: Postcolonial Differentials in Algorithmic Bias: Challenging Digital Neo-Colonialism in Africa. *SCRIPTed* 20, 194–226 (2023). <https://doi.org/10.2966/scrip.200223.194>
9. Birhane, A.: Algorithmic Colonization of Africa. *SCRIPTed* 17, 389–409 (2020). <https://doi.org/10.2966/scrip.170220.389>
10. Gentzel, M.: Biased Face Recognition Technology Used by Government: A Problem for Liberal Democracy. *Philosophy & Technology* 34, 1639–1663 (2021). <https://doi.org/10.1007/s13347-021-00478-z>
11. Priya, A., Arockiasamy, J.C.: The Global Reach of AI: A Postcolonial Analysis of Technological Dominance. *International Journal of Scientific Research in Science Technology* 11, 248–256 (2024).
12. Klo, S.T.: The Global South as Subversive Practice: Challenges and Potentials of a Heuristic Concept. *The Global South* 11, 21–37 (2017). <https://doi.org/10.2979/globalsouth.11.2.01>

13. Levander, C., Mignolo, W.: Introduction: The Global South and World DisOrder. *The Global South* 5, 11–11 (2011). <https://doi.org/10.2979/globalsouth.5.1.1>
14. Prashad, V.: *The Darker Nations: A People's History of the Third World*. 1st edn. New Press, New York (2008).
15. Reuters: DeepSeek narrows China-US AI gap to three months, AI founder Lee Kai-fu says. <https://www.reuters.com/technology/artificial-intelligence/deepseek-narrows-china-us-ai-gap-three-months-ai-founder-lee-kai-fu-says-2025-03-25>. Accessed 6 Aug 2025
16. Yeats, W.B.: *The Collected Poems of W. B. Yeats*. 1st edn. Macmillan, New York (1977).
17. Tran, H.: Africa: The Center of the Global South. Policy Brief PB-2432. Policy Center for the New South (2024).
18. Marwala, T.: *Heal Our World*. 1st edn. Pan Macmillan, Johannesburg (2022).
19. Furendal, M.: Collective Ownership of AI. In: Berkich, D., d'Alfonso, M.V. (eds.) *A Companion to Applied Philosophy of AI*, pp. 372–386. Wiley Blackwell, Bremen (2025).
20. Arnold, Z., Rahkovsky, I., Huang, T.: Tracking AI Investment. Center for Security and Emerging Technology Issue Brief (2020).
21. 36Kr: Investors flee AI. <https://36kr.com/p/1724425601025>. Accessed 8 Aug 2025
22. Morgan Stanley: AI in China: A Sleeping Giant Awakens. <https://www.morganstanley.com/insights/articles/china-ai-becoming-global-leader>. Accessed 8 Aug 2025
23. Calcalist: AI infrastructure boom: Tech giants spend 92 billion in a quarter. <https://www.calcalistech.com/ctech/news/article/yvyyf62gmr>. Accessed 5 Aug 2025
24. McKinsey Company: The state of AI in 2022—and a half decade in review. <https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai-in-2022-and-a-half-decade-in-review>. Accessed 29 July 2025
25. Buolamwini, J., Gebru, T.: Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. *Proceedings of Machine Learning Research* 81, 1–15 (2018).
26. White House: Ending Radical and Wasteful Government DEI Programs and Preferencing. <https://www.whitehouse.gov/presidential-actions/202501ending-radical-and-wasteful-government-dei-programs-and-preferencing>. Accessed 20 May 2025
27. Murray, C.: IBM Reportedly Walks Back Diversity Policies, Citing Inherent Tensions: Here Are All The Companies Rolling Back DEI Programs. *Forbes*. <https://www.forbes.com/sites/conormurray/20250322/mlb-removes-references-to-diversity-from-careers-website-here-are-all-the-companies-rolling-back-dei-programs>. Accessed 18 May 2025
28. Calvin, S., Leung, P.: Who owns artificial intelligence? A preliminary analysis of corporate intellectual property strategies and why they matter. Future of Humanity Institute Working Paper. <https://www.fhi.ox.ac.uk/wp-content/uploads/GovAI-working-paper-Who-owns-AI-Apr2020.pdf>. Accessed 7 Aug 2025
29. White House: Removing Barriers to American Leadership in Artificial Intelligence. <https://www.whitehouse.gov/presidential-actions/202501removing-barriers-to-american-leadership-in-artificial-intelligence>. Accessed 20 May 2025
30. Jakes, L., Mazzei, P.: At AI Summit in Paris, Vance Says U.S. Won't Sacrifice Innovation. *The New York Times*, 11 February 2025. <https://www.nytimes.com/2025/02/11/world/europe/vance-speech-paris-ai-summit.html>. Accessed 20 May 2025
31. Bishop, S.: Erotica, Gore and Racism: How America's War on Ideological Bias is Letting AI off the Leash. *The Conversation*. <https://theconversation.com/erotica-gore-and-racism-how-americas-war-on-ideological-bias-is-letting-ai-off-the-leash-250060>. Accessed 7 May 2025

32. Zuboff, S.: *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. PublicAffairs, New York (2019).
33. Deibert, R.J.: Three Painful Truths About Social Media. *Journal of Democracy* 30, 125–139 (2019). <https://doi.org/10.1353/jod.2019.0002>
34. Nemitz, P.: Constitutional Democracy and Technology in the Age of Artificial Intelligence. *Philosophical Transactions of the Royal Society A* 376, 20180089 (2018). <https://doi.org/10.1098/rsta.2018.0089>
35. Miller, K., Wolf, M., Grodzinsky, F.S.: Why We Should Have Seen That Coming. *ORBIT Journal* 1, 2–11 (2017).
36. Microsoft: Learning from Tay's Introduction. <https://blogs.microsoft.com/blog/2016/03/25/learning-tays-introduction>. Accessed 7 July 2025
37. Zemk, T.: Failure of Chatbot Tay was Evil, Ugliness and Uselessness in its Nature or Do We Judge it Through Cognitive Shortcuts and Biases? *AI Society* 36, 361–367 (2021). <https://doi.org/10.1007/s00146-020-01091-1>
38. Neff, G., Nagy, P.: Talking to Bots: Symbiotic Agency and the Case of Tay. *International Journal of Communication* 10, 4915–4931 (2016).
39. Grok: X post. <https://x.com/grok/status/1901475033975046320>. Accessed 20 May 2025
40. Vorsino, Z.: Chatbots, Gender, and Race on Web 2.0 Platforms: Tay.AI as Monstrous Femininity and Abject Whiteness. *Signs: Journal of Women in Culture and Society* 47, 189–215 (2021). <https://doi.org/10.1086/715238>
41. Ballim, Y., Breckenridge, K.: Divinatory Computation: Artificial Intelligence and the Future of the African Continent. Wits Institute for Social and Economic Research. <https://wiser.wits.ac.za/system/files/documents/BallimBreckenridgeAIfinal.pdf>. Accessed 12 Dec 2024
42. National Security Commission on Artificial Intelligence: The Final Report. <https://reports.nscai.gov/final-report>. Accessed 7 Aug 2025
43. Lee, K.F.: *AI Superpowers: China, Silicon Valley, and the New World Order*. 1st edn. Harper Business, New York (2018).
44. Gravett, W.: Digital Neo-Colonialism: The Chinese Model of Internet Sovereignty in Africa. *African Human Rights Law Journal* 20, 125–146 (2020). <https://doi.org/10.17159/1996-2096/2020/v20n1a5>
45. Kania, E.B., Webster, G.: China's AI Regulations and How They Get Made. Carnegie Endowment for International Peace. <https://carnegieendowment.org/research/2023/07/chinas-ai-regulations-and-how-they-get-made>. Accessed 8 July 2025
46. CB Insights: State of AI Q2 24 Report. <https://www.cbinsights.com/research/report/ai-trends-q2-2024>. Accessed 8 Aug 2025
47. Nkrumah, K.: *Neo-Colonialism: The Last Stage of Imperialism*. Thomas Nelson and Sons, Edinburgh (1965).
48. Munyua, A.W.: Exploring the Multi-Stakeholder Experience in Kenya. *Journal of Cyber Policy* 1, 206–221 (2016). <https://doi.org/10.1080/23738871.2016.1228990>
49. Calandro, E., Berglund, N.: African Internet Governance: Taking Stock and Looking Ahead. In: Calandro, E., Gillwald, A. (eds.) *Regional Informatics: ICT for Development Research and Policy in Africa*, pp. 23–45. Research ICT Africa, Cape Town (2019).
50. DeNardis, L.: *The Global War for Internet Governance*. Yale University Press, New Haven (2014).
51. Nyabola, N.: *Digital Democracy, Analogue Politics: How the Internet Era is Transforming Politics in Kenya*. Zed Books, London (2018).

52. Malcolm, J.: Multi-Stakeholder Governance and the Internet Governance Forum. Terminus Press, Perth (2008).
53. Quijano, A.: Coloniality of Power, Eurocentrism, and Latin America. *Nepantla: Views from South* 1, 533–580 (2000).
54. Rodney, W.: How Europe Underdeveloped Africa. Bogle-L’Ouverture Publications, London (1972).
55. Amin, S.: Unequal Development: An Essay on the Social Formations of Peripheral Capitalism. Monthly Review Press, New York (1976).
56. Fanon, F.: The Wretched of the Earth. Grove Press, New York (1961).
57. Thiongo, N.W.: Decolonising the Mind: The Politics of Language in African Literature. James Currey, London (1986).
58. Galtung, J.: Violence, Peace, and Peace Research. *Journal of Peace Research* 6, 167–191 (1969). <https://doi.org/10.1177/002234336900600301>
59. Porter, M.E.: Clusters and the New Economics of Competition. *Harvard Business Review* 76, 77–90 (1998).
60. Audretsch, D.B., Feldman, M.P.: Knowledge Spillovers and the Geography of Innovation. *Handbook of Regional and Urban Economics* 4, 713–739 (2004).
61. Selbst, A.D., Boyd, D., Friedler, S.A., et al.: Fairness and Abstraction in Sociotechnical Systems. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 59–68 (2019). <https://doi.org/10.1145/3287560.3287598>
62. Acemoglu, D., Robinson, J.A.: Why Nations Fail: The Origins of Power, Prosperity, and Poverty. Crown Publishers, New York (2012).
63. Friedler, S.A., Scheidegger, C., Venkatasubramanian, S., et al.: The Impossibility of Fairness: Different Value Systems Require Different Mechanisms for Fair Decision Making. *Communications of the ACM* 64, 136–143 (2021). <https://doi.org/10.1145/3433949>

“Where you want her, how you want her”

Understanding the violence of deepfake pornography

Amukelani Mnisi^{1[0000-0001-5443-8241]}

¹ University of the Witwatersrand, Johannesburg, Braamfontein 2000, South Africa
amnisi2000@icloud.com

Abstract. The growing popularity and adoption of generative artificial intelligence (AI) has given rise to the worrying phenomenon of deepfake pornography. A type of synthetic media generated by AI in which a person’s face is superimposed onto existing pornographic material, creating a new, hyper-realistic version in which a new person is depicted in pornographic content they did not actually participate in. Deepfake pornography is often created without the knowledge and consent of those depicted and is a form of technology-facilitated gender-based violence (TFGBV) that overwhelmingly targets women. There is a growing body of literature on the prevalence of deepfake pornography and other forms of TFGBV, however, little effort has been made to understand the purpose and function of this kind of violence against women. This research conceptualises a continuum of TFGBV to illustrate the range of behaviours that constitute TFGBV and identify the common character that underlies these behaviours. An analysis of three cases of deepfake pornography using this continuum reveals that deepfake pornography broadly functions to publicly violate women’s sexual autonomy with the purpose of reinforcing women’s domination and subjugation at the hands of men.

Keywords: Deepfake pornography, technology-facilitated gender-based violence, subjugation, domination.

1 Introduction

In January 2024, the topic “Taylor Swift AI” was trending on X. The global superstar had become the latest famous woman to be victimised by deepfake pornography. A type of synthetic media generated by artificial intelligence (AI) in which a person’s face is superimposed onto an existing pornographic image or video, creating a new, hyper-realistic version in which the new person is depicted performing in pornographic content they did not actually participate in [1, 2]. The fallout was widespread. Her fans flooded the social media platform with positive and supportive posts to overwhelm the pornographic AI-generated images. SAG-AFTRA, the powerful union representing media professionals in the US, condemned the images and vocalised support for legislation to prevent the creation of deepfake pornographic images. Even the Biden administration called for social media companies to play an active role in enforcing their

moderation rules to prevent the spread of non-consensual intimate imagery [1]. In response, X (formerly Twitter) wiped all search results of “Taylor Swift nude” and “Taylor Swift AI” from the platform [3].

While those pornographic deepfakes were widely condemned and the social media site quickly moved to erase the images, Swift’s experience greatly differs from the experiences of non-famous women who are victimised by deepfake pornography [4]. When Muslim Indian investigative journalist Rana Ayyub was victimised in 2018, the police attempted to shirk their responsibility and only filed her complaint after she threatened to go to the media with her story. Six months after her complaint, Ayyub had not received any communications or updates from the police on her case [5]. The average woman who discovers that she has become the victim of deepfake pornography can do little else other than submit take down requests to a social media platform that will go ignored [6]. For the many more women who are being victimised on imageboards and websites such as 4chan, they may never know that they are victims of deepfake pornography.

In recent decades, feminist scholars have extensively studied forms of image-based sexual abuse. In this context, deepfake pornography has quickly become seen as an act of technology-facilitated gender-based violence (TFGBV) [7]. Women are overwhelmingly targeted by deepfake pornography, with empirical research suggesting that over 95% of such content features women as the primary subjects [7, 8, 9]. Yet the purpose and function of this violence go largely unquestioned. This paper introduces a continuum of TFGBV as an analytical tool to begin answering the question of why women are being targeted in such a graphic and sexualised way. By placing deepfake pornography and other acts of TFGBV along a continuum, the common character of seemingly disparate acts of TFGBV can be identified and their purpose and function better understood.

Using this continuum to analyse three cases of deepfake pornography, this paper argues that the general function of deepfake pornography is two-fold. Firstly, it publicly violates women’s sexual autonomy. While its secondary function is the creation of an ambient culture of fear in which all women are reminded of their subordinate status and the ever-present threat of victimisation. This paper also argues that the purpose of this violence is to reinforce women’s domination, subjugation and objectification at the hands of men.

The remainder of this paper is organised into five sections. It begins with an explanation of the factors driving the proliferation and consumption of deepfake pornography across the internet and details the process of creating and distributing deepfake pornography on the imageboard 4chan. The three comparative case studies used in this research are also introduced and outlined. The literature review provides an overview of the growing body of research on deepfake pornography, locating this phenomenon in the debates around power, gender, and violence to which it is inextricably linked. It also provides the definitions and conceptual framework used to analyse the three cases

presented. A discussion of the methodology explains why and how a case study methodology was best suited to answer the research question. Analysis of the three case studies using the continuum of TFGBV illuminates how deepfake pornography both emerges from and works in tandem with other acts of TFGBV to target women with sexualised abuse online. The paper concludes with a discussion of the importance of viewing deepfake pornography and other forms of TFGBV as interconnected acts of violence that work to reinforce historical gender inequalities using new and evolving modes of perpetration.

2 Background

The ability to create deepfake pornography using generative AI first gained public attention in 2017 when a deepfake of actress Gal Gadot performing in an incest-themed porn video went viral [10, 11, 12]. It was posted on the online forum Reddit by a user named ‘deepfakes’, who created the video using freely available, open-source machine learning tools [12]. While this early deepfake was derided as unconvincing, it was still viewed as a worrisome harbinger of a growing phenomenon [12]. Since this first high-profile incident of deepfake pornography, the landscape for the creation, distribution and consumption of this content has changed in two significant ways.

The first change is that the technical quality of deepfake pornography has significantly improved. Most AI-generated images and videos are produced through a Generative Adversarial Network (GAN) machine learning model, which can generate new content based on existing data [13]. Inspired by the neural architecture and function of the human brain, GANs have two competing neural networks: a generator and a discriminator [13]. The generator creates realistic data samples while the discriminator distinguishes between real and generated samples, pushing the generator to improve its data generation capabilities until the discriminator can no longer differentiate between the two samples [13]. Initially, GANs were limited to small-sized images to maintain the model’s stability [14]. Since 2017, GAN models have improved with the introduction of the progressive learning training strategy, which expanded the capacity of the generator and discriminator, giving GANs the ability to generate large, high-quality images instead [14]. As a result, GANs can now generate high-resolution, photo-realistic images that, in turn, have made pornographic deepfakes look significantly more convincing [13].

The second change is that there are now far more people creating deepfake pornography across various platforms, as a subset of a broader proliferation of AI tools [15]. A McKinsey & Company survey on the adoption of AI in business, its impact, and investment in AI [16] found that from 2017 to 2022, the general use of AI had more than doubled. After the wildly successful launches of OpenAI’s AI tools in 2022 – one of which was DALL-E, an open source, publicly available AI image generator – big tech companies raced to create their own generative AI tools and integrate generative

models into their mainstream products. While these AI tools have safety guidelines and employ various misuse prevention techniques to detect and block users' requests to generate harmful content, they are easily circumvented, especially when the AI model is open-sourced [17]. A greater variety of generative AI tools to choose from and the accessibility and popular adoption of these tools have made it easier for users to generate deepfake pornography and as such, the volume of it has grown. Websites such as Mr DeepFakes, which was shut down in May 2025, hosted forty-three thousand deepfake pornographic videos as of November 2023 [18, 15]. On the website, there is a community of hundreds of deepfake content creators and sellers, with 657 users being responsible for over 95% of the site's video content, while other creators frequently respond to users' requests for custom deepfake content [15]. The content on Mr DeepFakes primarily featured celebrity women, with 95.3% of the individuals depicted being women musicians and actors [15].

However, it is not only famous women who are targets of deepfake pornography. On the infamous imageboard 4chan, the "Adult Requests" board has become a hotbed of deepfake pornography production, distribution, and consumption. On the board, anonymous users upload images of women and request a pornographic deepfake to be made of her. The request often includes some variation of the specific sex acts, positions, and number of sexual partners they want depicted in the deepfake. In some cases, requesters will provide an existing pornographic video they want the targeted woman to appear in. If other users are interested in the request, they will comment on the original request with a "Bump" to express interest and hopefully gain the attention of a user who knows how to create the desired deepfake [19]. Once the deepfake is created, the creator responds to the original post with a publicly accessible link through which the deepfake can be accessed by anyone, even those who are not 4chan users. It is not always clear how the user requesting the deepfake knows the woman whose photo they uploaded, as many of the threads are both anonymous, and the requests do not contain explanations. In many cases, these images were innocuous and unremarkable, the kind that anyone might post on their personal social media. They include photographs of women in everyday scenes, such as at their graduations, hanging out with their friends and families, or at the beach.

Whereas the "Adult Requests" board operates as a casual marketplace for the creation, distribution, and consumption of deepfake pornography ostensibly for personal use, other parts of 4chan have created and distributed deepfake pornography to serve as a weapon to be used against women. For several years beginning in 2014, feminist media critic Anita Sarkeesian became the target of a massive coordinated and vitriolic harassment campaign that originated on 4chan [20]. The harassment campaign was in response to a series of YouTube videos in which Sarkeesian critiqued outdated depictions of women in some video games [20]. To the gaming community and 4chan users – both of which are overwhelmingly male-dominated spaces due in part to the harassment women face when they attempt to enter these spaces – these videos were an attack on them and symbolised a larger feminist encroachment on an internet and gaming culture made for men and by men [21, 20]. One of the core features of the coordinated

harassment campaign was the creation and distribution of pornographic deepfakes of Sarkeesian that were shared across multiple platforms in a bid to harm her professional reputation [20]. Participants in the harassment campaign shared Sarkeesian's contact information online with the hopes of her being confronted with the video in her personal email accounts; it was their goal to make these pornographic deepfakes of her inescapable [22].

The online harassment of Anita Sarkeesian represents just one context in which deepfake pornography is used, but the three cases analysed in this paper show that deepfake pornography is weaponised against women for a myriad of reasons and to serve many different and context-specific purposes.

2.1 Case studies

Rana Ayyub

In April 2018, investigative journalist Rana Ayyub became the victim of what she called “a deepfake porn plot” intended to silence her [5]. Ayyub is one of India’s most prominent and globally recognised journalists and the author of the international best-seller *Gujarat Files: Anatomy of a Cover-Up* that exposed prime minister Narendra Modi’s role in the Gujarat riots of 2002 [23]. As a Muslim woman in India and vocal critic of the Modi government, Ayyub is frequently targeted with misogynistic and Islamophobic online abuse [5, 23]. After appearing on an Al Jazeera TV panel where she criticised India’s protection of child abusers after the rape of an eight-year-old girl, Ayyub received a pornographic deepfake video of herself from a source within India’s ruling Bharatiya Janata Party (BJP) [5]. The deepfake spread across online chat groups maintained by BJP supporters, went viral, and was even shared on the Facebook page of the chief minister of India’s largest state [23]. Screenshots from the video were posted in the comments section of Ayyub’s social media posts and shared with her family members [5].

Ayyub faced a deluge of sexual harassment across various social media platforms, with people messaging her to enquire about her “rates for sex” and some threatening her with gang rape if she continued to criticise the Modi government [5]. The next day, Ayyub was doxxed with her home address and personal phone number being posted publicly on social media [5, 23].

After an international outcry, which included the United Nations writing to the Indian government, the abuse slowed [24]. However, Ayyub continues to feel its impact. She lives in fear that she could be targeted like this again and, as such, uses social media more cautiously than before [5]. Despite the caution, Ayyub still experiences waves of online abuse, particularly after the publication of any of her reporting or commentary

on communal violence, human rights violations, and politics in India [23]. The abuse is misogynistic, often sexually explicit, and seeks to undermine her credibility as a journalist [23].

Taylor Swift

In January 2024, pornographic deepfake images of Taylor Swift circulated across social media, accumulating tens of millions of views [1, 4]. The images depicted Swift in several sexualised positions and partially clothed in Kansas City Chiefs merchandise, the National Football League (NFL) team her boyfriend Travis Kelce plays for.

The origin of the deepfake images was traced back to a community on the imageboard 4chan, which encouraged users to evade the safety policies of AI image generator tools by attempting to create pornographic and sometimes violent images of famous women [25]. It became something of a game on the imageboard with users being encouraged to “be creative,” share tips for generating the pornographic images, and commending each other for good work [25]. Having first appeared in this 4chan community on January 6, 2024, the images appeared eleven days later on the encrypted messaging app Telegram – in a group dedicated to abusive images of women – before appearing on mainstream social media a day later [25].

As an incredibly successful woman in the public eye, Swift has experienced misogyny throughout her career [26]. Critics view her as undeserving of her success, are dismissive of her talents, and express a deep disdain for her public persona [26]. Swift’s public persona is notable for how clearly it is designed to appeal to women and girls, rather than men, and the intense devotion this persona has inspired amongst her global fanbase. [27]. This persona, and the fierce support it engenders amongst Swift’s fanbase of predominantly young women, has been a frequent target of misogynistic derision [26].

However, once Swift’s relationship with Kelce became public knowledge, there was a noticeable increase in the hostile sexism she experienced online [27]. When Swift began attending Kelce’s football games, she bought with her the intense media interest she generates and her large fanbase, which precipitated a 24% increase in women aged 18-24 who watched the 2024 Super Bowl [28]. Where misogynistic rhetoric against Swift and her fans was once largely confined to dedicated online anti-fan communities, it was now also coming from the NFL’s large, and largely male, fanbase as well [27].

While Swift never publicly commented on the pornographic deepfakes made of her, they drew widespread ire and were viewed by some as an attempt by men to put a successful and powerful woman “back in her box” and remind her that no matter how

powerful she may be, men can still “reduce you to a sex object and there’s nothing you can do about it” [4].

4chan

Since its creation in 2003, 4chan has become infamous in popular culture for its incredibly lenient content moderation policies, controversial subcultures, and alleged hosting of illegal and unethical activity, including deepfake pornography [9, 29]. As previously described, the “Adult Requests” board on 4chan is a hotbed of deepfake pornography creation, distribution, and consumption. While the board is often also used by users attempting to find the names of porn performers and OnlyFans content creators, most requests on the board are requests for the creation of pornographic deepfakes.

The board operates as a kind of impersonal online marketplace. Both the users who upload images of women requesting pornographic deepfakes and the users who create them are anonymous, and no money or exchange for the ‘service’ of creating the deepfake takes place. Compared to other parts of 4chan known for their offensive language and hostile communication, the “Adult Requests” board is relatively tame, and discussion rarely goes beyond users specifying how they want the women depicted in the deepfake (i.e.: topless, legs spread, etc.) and later thanking the user who creates the desired deepfake. On the “Adult Requests” board, the objectification of women is simply routine.

On one thread on the board, dated April 14, 2025, an anonymous user offers their services to others, posting on the board, “I’m making deepfakes for the next 2 hours. Post your brown or black-haired crushes” [30]. All the user asks for is to be told “how you know her, where you want her, how you want her.” The post received numerous responses with one user uploading an image of a woman and expressing a desire to “see this broad’s titties” [30]. Another user posted an image of a woman identified as a former coworker, and a request to see her depicted in a deepfake with her legs spread; the user provided a reference image for the desired deepfake [30]. With no fanfare, the original poster provides the other users with their requested deepfakes, posting the links on the publicly available thread.

Whether women are highly visible celebrities, posting innocuous images of themselves online, or simply doing their jobs, the violence of being targeted with deepfake pornography remains ever-present for all women. As AI tools grow more popular and sophisticated with each year and the volume of deepfake pornography online grows, greater awareness and research of this phenomenon becomes increasingly necessary and urgent.

3 Literature review

The problem of deepfake pornography is inextricably linked to issues of gender, violence, and power. Therefore, furthering the existing research on deepfake pornography and understanding the function and purpose of this violence requires this research to grapple with violence against women and its evolving modes of perpetration, pornography, and the pervasiveness of patriarchal violence in all these issues. The literature reviewed reflects the layers and often interconnected nature of these issues in trying to make sense of deepfake pornography.

The first body of literature this research engages with is violence against women and its changing modes of perpetration. In the growing body of research on deepfake pornography, it is classified as a form of technology-facilitated gender-based violence (TFGBV) generally and an act of image-based sexual abuse specifically [7]. TFGBV is a form of gender-based violence that, like other forms of gender-based violence, is rooted in the systemic reinforcement of gender norms and inequality and uses technology to broaden the scope of violence it inflicts on its victims, primarily women and girls [7]. The use of technology has introduced new forms of abuse, such as ‘revenge porn’ and the creation of deepfake pornography [31]. Acts of TFGBV are considered acts of technology-facilitated violence because, like other technology-facilitated attacks, they have the ability to dominate, intimidate, dehumanise, and create reasonable feelings of fear in those targeted by them [7, 32, 11].

While pornographic deepfakes are the product of a technologically advanced epoch, they are also a continuation of historical patterns of sexual violence that overwhelmingly victimise women. Therefore, it can be placed on the continuum of sexual violence against women theorised by Liz Kelly [33]. Kelly uses the concept of a continuum in two ways. The first is to allow for a general discussion of sexual violence and identify the common character that underlies the different forms of violence as the “abuse, intimidation, coercion, intrusion, threat, and force men use to control women” [33]. The second is to use the continuum to document and name the range of abuse, intimidation, coercion, intrusion, threat, and force men use to control women while acknowledging that there are no defined and distinct analytic categories into which men’s behaviour can be categorised [33]. In this case, the range of the continuum does not denote severity but rather the prevalence of certain acts of sexual violence [33]. Therefore, acts of sexual violence, such as sexual harassment, that most women experience in their lives and are likely to experience on multiple occasions due to these acts being viewed as acceptable male behaviour, are placed at the beginning of the continuum [33]. Other acts of sexual violence, such as rape, which are defined as criminal, appear at the end of the continuum. The continuum does not suggest that there is a linear progression connecting different acts of sexual violence, but does illuminate how common and accepted forms of sexual violence lay the foundation for the perpetration of other, sometimes criminal, forms of sexual violence [33].

McGlynn et al. [34] build on Kelly's concept of the continuum in the context of ubiquitous technology. They argue that while image-based sexual abuse (IBSA) can and should be placed on Kelly's originally theorised continuum, our shared concept of IBSA itself is unique to a time in history in which technology is an integral part of our everyday lives and offers new modes of perpetration of violence against women [33]. Thus, on the continuum formed by practices and behaviours of image-based sexual abuse, the common character of the various forms of IBSA are: the sexual nature of the imagery; the gendered nature of the abuse; the sexualised nature of the harassment and abuse; the harms as breaches of fundamental rights to dignity, sexual autonomy and sexual expression; and the minimisation of these forms of abuse [34]. Though the two continuums were conceptualised in vastly different socio-technological landscapes, they are united in analysing the harms created by behaviours on their respective continuums, including 'abuse, intimidation, intrusion, threat and force' [33].

From the existing literature on image-based sexual abuse comes an expanded definition of violence that focuses not only on physical violence and its impact but also foregrounds the cumulative impact of psychological attacks [7]. This violence, whether physical and psychological, is rooted in a reality significantly shaped by existing gendered and racialised systems of power. The continuums of sexual violence and IBSA theorised by Kelly [33] and McGlynn et al. [34], respectively, are helpful conceptual tools that ground and locate deepfake pornography in a long history of violence against women and its different modes of perpetration.

In the context of deepfake pornography and the violence of creating, distributing, and consuming it, this research must contend with a second body of literature on the politics of pornography. For Dworkin [35], pornography reinforces and reflects the widespread and socially sanctioned objectification and debasement of women. Moreover, the legitimisation of pornography as "depictions of the erotic" communicates to the viewer that the real pleasure of sex is the debasement of women [35]. While Dworkin's arguments represent only one side of the feminist debate about pornography, in the case of deepfake pornography they are invigorated with renewed relevance because of how this kind of pornography acts as a violation of women's autonomy. Since Dworkin made these arguments, pornography has only grown more violent and more widely accessible. This primes generations of young men who watch pornography to find sexual satisfaction and masculine affirmation in seeing women be sexually humiliated [36].

In many ways, Dworkin already observed and understood that the pornography industry would only continue to grow and with it, the need for more bodies and the sexual brutalisation of those bodies. Dworkin knew that "The numbers [of pictures needed to meet the demand for pornography] grow as the technology and its accessibility grow" [35]. Deepfake pornography is the result of this unrelenting demand, and this research question is an effort to understand the violence that ensues when the demand never ceases, and the technology and accessibility only continue to grow.

In the existing research on deepfake pornography, McGlynn et. al [34] have come the closest to making sense of the function and purpose of its violence, albeit solely from the perspective and experiences of victim-survivors. This paper and the methodology employed to answer this question attempt to contextualise incidents and processes of producing, distributing, and consuming deepfake pornography and, in doing so, come closer to understanding the motivations of the perpetrators and the effects of these acts of TFGBV.

3.1 Conceptual framework

A continuum is specifically designed to illustrate range, and when applied to sexual violence against women and the behaviours that constitute image-based sexual abuse, it allows for a general discussion about these types of violence and the discovery of the common character that underlies a range of violent behaviours. Thus, to understand how we arrive at the violence of creating, distributing, and consuming deepfake pornography, a continuum is perfectly suited to present the range of TFGBV and the common character that underlies different acts of TFGBV. Moreover, a continuum demonstrates the relationship between technology, gender, and power – the concepts at the core of this research – and how they operate in tandem to victimise women online.

Unlike the originally theorised continuum of sexual violence, the continuum of TFGBV conceptualised in this research is organised according to severity, as some of the acts listed have the potential to cause serious, offline harms for victims, whether they be physical or psychological.

The acts of TFGBV that make up the first portion of the continuum can be broadly categorised as ‘e-bile.’ E-bile is hostile misogynistic rhetoric laced with profanity and graphic threats of sexual violence [37]. While it is often deeply disturbing to read, e-bile has become a standard discursive tool when internet users want to express their disagreement with or disapproval of women [37]. It is rooted in hostile sexism – negative attitudes and contempt for women, specifically those who are perceived as transgressing traditional gender roles – and is most often directed at women who are highly visible and outspoken in the public sphere, such as woman politicians, although all women can and are targeted as well [37, 27]. Closely related to e-bile, although it should be considered as a separate act of TFGBV, is ‘lascivious contempt’. Lascivious contempt targets women who are perceived as hypersexualised and then derogated as ‘sluts’ who are unsatisfactory because they are too fat, too ugly, etc. [37]. What makes lascivious contempt different from e-bile is that it makes a competition of misogynist rhetoric, usually playing out in social media comment sections where users compete to make the most offensive comment, often escalating to increasingly sexualised and graphic insults [37].

The midpoint of the continuum of TFGBV is when online behaviour begins to create offline harms such as negatively impacting the targeted person's professional reputation and threatening their current and future employment. The 'Gamergate' coordinated harassment campaign against Anita Sarkeesian is an example of this. The impact of the campaign was felt not only online in the deluge of death and rape threats Sarkeesian received, but also offline. A speech that Sarkeesian was scheduled to deliver at Utah State University was cancelled after the university received an email threatening "a massacre-style attack" if the talk proceeded and a bomb threat was called in at the Game Developer's Choice Awards where Sarkeesian was set to receive an award [38, 39]. These hate campaigns often emerge and escalate in the face of feminist activism and perceived feminist gains in spaces traditionally geared towards men (Jane, 2014). This was the case for Sarkeesian, with a member of the gamer community celebrating her being targeted with deepfake pornography because "she attacked us first" by attempting to critique portrayals of women in some video games [22].

With its overlapping presence in coordinated harassment campaigns, it could be argued that deepfake pornography and other forms of image-based sexual abuse (IBSA) are the next addition to the continuum of TFGBV. In the case of Rana Ayyub, the distribution of deepfake pornographic images of her operated as a coordinated harassment campaign with the impacts being felt both on and offline [5, 23]. These acts of IBSA, which include revenge porn and the non-consensual distribution of intimate images, may be directed at one specific woman at any given moment, but they serve as a reminder to all women that it could happen to them. For victims of IBSA, it creates a near-constant fear that these images will be discovered by family members, future employers, or partners, and thereby negatively affect their futures offline [11, 4].

The furthest end of the continuum of TFGBV would be 'Doxxing,' because of its potential to cause serious physical harm to its targets. Doxxing is the non-consensual publication of another person's private or personal identifying information – such as their phone number and home address – on the internet, often to enable their online harassment [40, 20]. Doxxing is often a feature of coordinated harassment campaigns and is compounded by the other features of harassment campaigns like e-bile, however, it is placed at the furthest end of the continuum for three reasons. Firstly, it gives those who are making violent threats against others on the internet the information needed to actualise those threats and cause real physical harm if they choose [40]. Secondly, once that private information is published, it can be diffused across the internet which may invite further harm against the person doxxed long after its initial publication [40]. Finally, for victims of doxxing, they are aware that because of the spread of their information, there is no foreseeable endpoint to the potential danger being doxxed has put them in [40]. For victims, the harm of being doxxed is not experienced as a discrete event but as a repetitive one, an experience which can become a great source of psychological anguish for them [40].

The most cursory examination of the three cases used in this research already reveals the presence of other forms of TFGBV, such as e-bile and coordinated harassment campaigns, in instances where women were targeted with deepfake pornography. That women are being targeted with the same acts of TFGBV suggests that these acts share a common character and are being deployed to serve similar, if not the same, function and purpose. While the continuum of TFGBV comprises seemingly distinct acts of violence, in practice, these acts of violence are combined or escalate to create enormous harm for victims and targets. Without the continuum of TFGBV, deepfake pornography is decontextualized and treated as separate from the behaviours that created and enabled it.

Therefore, a continuum of TFGBV illuminates the range of violence women experience online, as well as how these acts of violence seamlessly shade into each other as they progress along the continuum. Moreover, by placing deepfake pornography on the continuum of TFGBV alongside other acts of violence, we can identify the common character of these acts of violence and ultimately understand the function and purpose of the violence of deepfake pornography.

4 Methodology

This research seeks to understand the function and purpose of the violence of deepfake pornography. This requires understanding the motivations of creators, distributors, and consumers of deepfake pornography and its effects on them, as well as the women targeted. This research uses a case study methodology because it provides the opportunity for multi-perspectival analysis [41]. A case study model also offers the best opportunity to analyse and understand these differing perspectives and to examine them from a macro, societal level.

This research makes use of three case studies to maximise how much can be learnt about the function and purpose of the violence of deepfake pornography. The case studies selected for this research were chosen because they illustrate different contexts in which deepfake pornography was created and allow for reasonable inferences to be made about the different functions and purposes of the creation, distribution, and consumption of deepfake pornography in different contexts.

Each case required different units of observation and offered different sources of information. For Rana Ayyub, this research was able to draw from Ayyub's first-hand account of the events surrounding the deepfake pornography campaign against her and understand its direct effect on her as a victim of this kind of violence [5]. For Taylor Swift, Swift did not make any public statements about the viral deepfakes of herself and thus this research relied on the societal reaction to the deepfakes as it played out on social media. In the case of deepfake pornography creation, consumption, and distribu-

tion on 4chan, I monitored the “Adult Requests” board for a week to witness and understand the process and sought out older posts on the board to see if there were any changes in the process or in the behaviour and language of users.

While each unit of observation and source illuminated the details and complexities of each case, they weren’t without their own limitations. The use of Rana Ayyub’s first-hand account, while valuable as a primary source, by its nature focuses on one woman’s experience of being targeted with deepfake pornography and cannot and should not be used as representative of all victims’ experiences. For the Taylor Swift case, it is impossible to capture each reaction to the pornographic deepfakes and to keep track of the discussions it sparked, especially in such a fragmented online space like X (formerly Twitter). Similarly, a week of following posts on 4chan could not capture every user interaction on the board, nor could it keep pace with the sheer number of posts being created throughout that week.

Despite these challenges and limitations, a multiple case study methodology is the best approach to answering this research question because it highlights the similarities in seemingly disparate cases. The different sources that came with each case provided the multi-perspectival analysis needed to understand the phenomenon of deepfake pornography not just from the perspective of those involved in it (i.e.: the creators, consumers and victims) but from a broader society who is now aware, perhaps for the first time, of the existence of this form of violence and its implications for their own online safety.

5 Analysis and Discussion

In each of the cases studied in this research, deepfake pornography served a specific and unique function and purpose. An analysis and understanding of these specificities and differences helps us understand the broader function and purpose of the violence of this phenomenon.

For Rana Ayyub, the function of the deepfake pornography campaign against her was to silence and undermine her as a critic of the Modi government. When Ayyub’s case is placed on the continuum of technology-facilitated violence (TFGBV), the distinction between different acts of TFGBV disappears almost entirely. Ayyub became the target of increased misogynistic and Islamophobic e-bile, a harassment campaign with deepfake pornography at its centre, and was later doxxed by social media users. These acts of TFGBV were simultaneously deployed against Ayyub to form a multi-pronged assault that is increasingly used against women journalists around the world to shame, silence, undermine, and discredit them professionally [42]. While Ayyub had grown used to e-bile as an outspoken critic, the experience of being targeted by deepfake pornography impacted her so deeply that she remains cautious of what she posts online and self-censors as a means of protecting herself from being targeted again [5].

When Taylor Swift was targeted with deepfake pornography, those deepfakes functioned to reassert masculine dominance in the face of perceived feminist encroachment on male spaces. Taylor Swift's public persona, success, and deeply devoted fanbase had long attracted misogynistic derision and hostile sexism from dedicated anti-fan communities online [26]. However, that hostile sexism had never before prompted an act of violence such as the creation and distribution of pornographic deepfakes until Swift began dating NFL player Travis Kelce. When Kelce and Swift's relationship became public knowledge and Swift began attending his games, she bought her fanbase along with her and precipitated a 24% increase in women aged 18-24 who watched the 2024 Super Bowl [28]. The increased presence and participation of young women in the male-dominated football fandom was also compounded by the subversion of traditional gender norms that Swift's presence represented [27]. Despite playing the role of devoted girlfriend to a decorated athlete, Swift's status as a billionaire cultural juggernaut positions her as the proverbial breadwinner in her relationship and in violation of the masculinist ideals found amongst the NFL's fanbase [27]. Thus, when Swift was targeted with pornographic deepfakes, it read to women online as an attempt to place a powerful woman "back in her box" and reduce her to little more than a sex object [4]. Moreover, to have Swift wearing the merchandise of her boyfriend's football team in those deepfakes could be read as an attempt to define Swift and her sexuality in relation to the man in her life, asserting his masculine dominance over her and challenging the independence she enjoys in reality. While the use of pornographic deepfakes was specifically designed and targeted to negatively affect Rana Ayyub, the pornographic deepfakes of Taylor Swift affected a much larger audience. It is impossible to know how the deepfakes impacted Swift – she never made a public statement – but for the women online who saw them, it was a frightening demonstration of the unique violence women face online and emphasised the ever-present threat of victimisation for all women.

For both Rana Ayyub and Taylor Swift, the weaponisation of pornographic deepfakes against them was spurred by external events. Both women were frequent targets of hostile sexism and had experienced other acts of TFGBV such as e-bile, but the creation and distribution of pornographic deepfakes were an escalation in response to other events. For Ayyub, it was speaking out against sexual violence against women and children in India and her public perception amongst Hindu nationalists as a hired anti-Hindu and anti-Modi mouthpiece [23]. For Swift, it was her highly publicised and influential entrance into the male-dominated world of the NFL [27]. In both cases, the women's behaviour and presence were deemed as transgressing traditional gender norms and threats to patriarchy that had to be neutralised. Where other acts of TFGBV had failed to curb the women's success and outspokenness, deepfake pornography was deployed as a new and deeply intimate act of violence that would not only directly impact these women but would send a clear message to all women.

Unlike the other two cases, the creation, distribution, and consumption of deepfake pornography on 4chan does not represent an escalation of violence but rather an evolution of a prevailing culture of misogyny. On 4chan, a standard response to users finding

out that a woman is participating in a discussion or another user claims to be a woman is to demand that she show ‘tits or get the fuck out’ [20]. On 4chan, these acts of TFGBV are normalised and are merely the cost of entry. As previously described, the “Adult Requests” board is largely free from the hostile communication that characterises other parts of the infamous website. When there is brief discussion on the board, as the process of creating and distributing deepfakes takes place, users often employ the violent and degrading language that is frequently used in pornography. Women are described as “sluts” and requests are made for the women to be depicted in deepfakes enduring violent sex acts like “face fucking” [19]. This language and the creation of deepfake pornography are not out of place on a website that was the epicentre of the harassment campaign against Anita Sarkeesian and was the origin point of the pornographic deepfakes of Taylor Swift [25]. What is different about the process taking place on the “Adult Requests” board is that it appears almost purposeless. It has become mundane for users on the board to strip women of their autonomy and treat them as publicly available objects whose identities and likeness can be hijacked to fulfill any manner of sexual fantasies. In the case of 4chan, deepfake pornography functions to normalise women’s sexual objectification.

Despite the differences in each case, when placed on the continuum of TFGBV, it becomes clear that deepfake pornography is an act of violence that often works in tandem with other acts of TFGBV. Moreover, in each instance in which it is used, common characteristics emerge: the sexual nature of the imagery used to target women; the gendered nature of the abuse; the sexualised nature of the harassment and abuse; the harms as breaches of fundamental rights to dignity, sexual autonomy, and sexual expression [34]. Whether deepfake pornography is being used to silence and undermine political critics, reassert masculine dominance, or reinforce the mundanity and inevitability of misogyny, it does so by relying on sexualised imagery of women designed to reflect their objectification and subjugation at the hands of men and uses a mode of perpetration that leaves all women vulnerable to victimisation. In these cases, we see that the broad function of deepfake pornography is to sexually humiliate women and publicly violate their sexual autonomy with the purpose of reinforcing women’s domination and subjugation at the hands of men.

6 Conclusion

The range of available generative AI tools and their growing popularity have led to a significant increase in the volume of deepfake pornography online. While women in the public eye are often the most at risk of being targeted, all women are at risk of victimisation. The purpose of deepfake pornography is to create an ambient fear of victimisation in all women and thus reinforce their subordination and objectification at the hands of men. Deepfake pornography serves this purpose by functioning as an act of sexual humiliation and violation by depicting women – particularly women whose success or outspokenness challenge traditional gender roles – in pornographic content

that reinforces their role as sexual objects and vessels of male sexual desires and aggression.

The cases of Rana Ayyub, Taylor Swift, and the “Adult Requests” board on 4chan represent the different contexts in which deepfake pornography is used and weaponised. Analysis of these cases using the continuum of TFGBV reveals how deepfake pornography and other acts of TFGBV are often simultaneously deployed against women online because of their common character and goal of breaching women’s fundamental rights to dignity, sexual autonomy, and sexual expression. The continuum also highlights the importance of viewing deepfake pornography and other forms of TFGBV as interconnected and evolving acts of violence that are used to terrorise women online and reinforce the inevitability of their victimisation as a means of fortifying entrenched gender inequalities online.

References

1. Hibberd, J., Phillips, Z.G. (2024) Taylor Swift AI Nudes Provoke Fandom Uproar on X: “Disgusting as Hell”. The Hollywood Reporter. <https://www.hollywoodreporter.com/news/general-news/taylor-swift-ai-nudes-twitter-x-1235806979>, last accessed 2025/05/29.
2. Öhman, C: Introducing the pervert’s dilemma: a contribution to the critique of Deepfake Pornography. *Ethics and Information Technology*, 133-140 (2019).
3. Piña, C. (2024) Taylor Swift Search Re-enabled on X Following AI Nudes. The Hollywood Reporter. <https://www.hollywoodreporter.com/news/general-news/taylor-swift-search-re-enabled-x-ai-nudes-1235810070>, last accessed 2025/03/17.
4. Saner, E. (2024) Inside the Taylor Swift deepfake scandal: ‘It’s men telling a powerful woman to get back in her box.’ *The Guardian*. <https://www.theguardian.com/technology/2024/jan/31/inside-the-taylor-swift-deepfake-scandal-its-men-telling-a-powerful-woman-to-get-back-in-her-box>, last accessed 2025/03/21.
5. Ayyub, R (2018) I Was The Victim Of A Deepfake Porn Plot Intended To Silence Me: Rana Ayyub. *HuffPost*. https://www.huffpost.com/archive/in/entry/deepfake-porn_in_5c1201cfe4b0508b213746bd, last accessed 2025/02/15.
6. Human Rights Watch (2021) “My Life is Not Your Porn”: Digital Sex Crimes in South Korea. Human Rights Watch, New York.
7. Dunn, S.: Technology-Facilitated Gender-Based Violence: An Overview. Centre for International Governance Innovation, Waterloo (2020)
8. Security Hero (2024) State of Deepfakes: Realities, Threats, and Impacts. <https://www.securityhero.io/state-of-deepfakes/>, last accessed 2025/03/20.
9. Deeptrace (2019) The State of Deepfakes: Landscape, Threats, and Impact. https://regmedia.co.uk/2019/10/08/deepfake_report.pdf, last accessed 2025/04/20.
10. Wagner, T.L., Blewer, A.: “The Word Real Is No Longer Real”: Deepfakes, Gender, and the Challenges of AI-Altered Video. *Open Information Science* 2019(3), 32 – 46 (2019) <https://doi.org/10.1515/opis-2019-0003>.
11. Brieger, A. R.: Empowerment or exploitation: A qualitative analysis of online feminist communities’ discussions of deepfake pornography. Two-year Master’s Thesis. Uppsala Universitet (2024).
12. Cole, S. (2017) AI-Assisted Fake Porn Is Here and We’re All Fucked. *Vice*. <https://www.vice.com/en/article/gal-gadot-fake-ai-porn>, last accessed 2025/02/27
13. Banh, L., Strobel, G: Generative artificial intelligence. *Electronic Markets*, 62-79 (2023)
14. Li, C., Zhuang, B., Wang, G., Liang, X., Chang, X., Yang, Y. (2024) Automated Progressive Learning for Efficient Training of Vision Transformers. Paper presented at Conference on Computer Vision and Pattern Recognition, Seattle, United States, 2024, <https://doi.org/10.48550/arXiv.2410.00350>.
15. Han, C., Li, A., Kumar, D., Durumeric, Z.: Characterizing the MrDeepFakes sexual deepfake marketplace. In: Proceedings of the 34th USENIX Conference on Security Symposium, pp. 5169–5188. USENIX Association, Seattle (2025).
16. McKinsey & Company. (2024) What is generative AI? <https://www.mckinsey.com/featured-insights/mckinsey-explainers/what-is-generative-ai>, last accessed 2025/04/15.
17. Clifford, B. (2023) Preventing AI Misuse: Current Techniques. Centre for the Governance of AI. <https://www.governance.ai/analysis/preventing-ai-misuse-current-techniques>, last accessed 2025/04/23.

18. Wise, A. (2025) Major deepfake porn site shuts down. NPR. <https://www.npr.org/2025/05/06/nx-s1-5388422/mr-deepfakes-porn-site-ai-shut-down>, last accessed 2025/10/3.
19. 4chan. n.d.
20. Nagle, A: Kill All Normies: Online Culture Wars from 4chan and Tumblr to Trump and the Alt-Right. Zero Books: United Kingdom (2017).
21. Andersen, I. V.: Hostility online: Flaming, trolling, and the public debate. First Monday 26(3). <http://dx.doi.org/10.5210/fm.v26i3.11547> (2021)
22. Harwell, D. (2018) Fake-porn videos are being weaponized to harass and humiliate women: ‘Everybody is a potential target’. The Washington Post. <https://www.washingtonpost.com/technology/2018/12/30/fake-porn-videos-are-being-weaponized-harass-humiliate-women-everybody-is-potential-target/>, last accessed 2025/03/21
23. Posetti, J., Bontcheva, K., Zaffar, H., Shabbir, N., Maynard, D., Pandya, M.: Rana Ayyub: Targeted online violence at the intersection of misogyny and Islamophobia. International Center for Journalists: Washington DC (2023).
24. United Nations (2018) UN experts call on India to protect journalist Rana Ayyub from online hate campaign. United Nations Office of the High Commissioner for Human Rights. <https://www.ohchr.org/en/press-releases/2018/05/un-experts-call-india-protect-journalist-rana-ayyub-online-hate-campaign>, last accessed 2025/03/10.
25. Hsu, T. (2024) Fake and Explicit Images of Taylor Swift Started on 4chan, Study Says. The New York Times. <https://www.nytimes.com/2024/02/05/business/media/taylor-swift-ai-fake-images.html>, last accessed 2025/03/25.
26. Kinnunen, L. (2025) In: Aker, P., Kaun, A. (eds) Gendering Media: Framing of AI, Interacting with ChatGPT and Anti-fandom. Södertörn University: Stockholm.
27. Elder, L., Greene, S., Gulatti, J., Lizotte, M. (2025) The Politics of Taylor Swift: Partisanship, Gen Z, and Sexism. Paper presented at Southern Political Science Association Conference 2025, San Juan, Puerto Rico.
28. Feldman, J. (2024) One Year Later, Taylor Swift Effect Stretches Far Beyond NFL. Sportico. <https://www.sportico.com/personalities/people/2024/taylor-swift-effect-nfl-attendance-girls-social-marketing-1234798352/>, last accessed 2025/05/1.
29. Satter, R. (2025) Notorious messageboard 4chan has been hacked, posts claim.” Reuters. <https://www.reuters.com/technology/cybersecurity/notorious-internet-messageboard-4chan-has-been-hacked-posts-claim-2025-04-15/#:~:text=It%20has%20long%20been%20notorious,nonconsensual%20pornography%20%2C%20opens%20new%20tab...>, last accessed 2025/04/16.
30. Anonymous. (2025) “Making Deepfakes of your brunettes” [4chan post]. 4chan.
31. Dunn, S. (2020) Is it Actually Violence? Framing Technology-Facilitated Abuse as Violence. In: Bailey, J.F. (ed) The Emerald International Handbook of Technology-Facilitated Violence and Abuse. Emerald Publishing Limited: Leeds.
32. McGlynn, C., Johnson, K., Rackley, E., Henry, N., Gavey, N., Flynn, A., Powell, A. ‘It’s Torture for the Soul’: The Harms of Image-Based Sexual Abuse. Social & Legal Studies, 541-562 (2021).
33. Kelly, L.: Surviving Sexual Violence. University of Minnesota Press, Minneapolis (1988).
34. McGlynn, C., Rackley, E., Houghton, R.: Beyond ‘Revenge Porn’: The Continuum of Image-Based Sexual Abuse. Feminist Legal Studies, 25-46 (2017).
35. Dworkin, A.: Pornography: Men Possessing Women. Penguin, New York (1981).
36. DeKeseredy, W. S., Schwartz, M. D.: Thinking Sociologically About Image-Based Sexual Abuse: The Contribution of Male Peer Support Theory. Sexualization Media & Society, 1-8 (2016).

37. Jane, E.A.: ‘Back to the kitchen, cunt’: speaking the unspeakable about online misogyny. *Continuum: Journal of Media and Cultural Studies* 28(4), 558-570 (2014)
<http://dx.doi.org/10.1080/10304312.2014.924479>.
38. Kelion, L. (2014) Feminist video-games talk cancelled after massacre threat. BBC.
<https://www.bbc.com/news/technology-29626809>, last accessed 2025/05/25
39. Seppala, T.J. (2014) GDC award for culture critic Anita Sarkeesian led to bomb threat. Engadget. <https://www.engadget.com/2014-09-17-anita-sarkeesian-gdc14-bomb-threat.html>, last accessed 2025/05/30
40. Anderson, B., Wood, M.A.: Harm Imbrication and Virtualised Violence: Reconceptualising the Harms of Doxxing. *International Journal for Crime, Justice and Social Democracy*, 11(1), 196-209 (2022).
41. Tellis, W.M.: Application of a Case Study Methodology. *The Qualitative Report* 3(3) (1997).
42. International Center for Journalists. (2022). *The Chilling: A global study of online violence against women journalists*. International Center for Journalists: Washington DC.

An initial foray into AI judicial drafting: a comparative experiment in a constitutional privacy case

Kudakwashe Mawowa¹ [0009-0002-5880-838X], Donrich Willem Thaldar¹ [0000-0002-7346-3490]

1 School of Law, University of KwaZulu-Natal, Durban, South Africa

Abstract. This empirical study investigates how well can Artificial Intelligence (AI) draft a South African High Court judgment using OpenAI's ChatGPT-4o large language model and case data from *De Jager v Netcare Limited and Others* 2025 JDR 0793 (GP)—an informational privacy and data protection dispute. Drawing from local and international experiments, the study develops a value-driven prompting framework and distills a rubric for evaluating the quality of AI-driven judgments. Twenty sequential prompts were developed to generate an artificial judgment which was then compared with the human decision with respect to structure, factual accuracy, use of legal authority, legal reasoning and clarity. ChatGPT-4o led across all five metrics, though formatting inadequacies were a weakness. This study did not assess the substantive legal correctness of either judgment. However, it is interesting that both judgments reached nearly identical outcomes on the main substantive legal issues, viz privacy, statutory compliance (Protection of Personal Information Act 4 of 2013) and admissibility of surveillance data. This may be attributable to their common reliance on similar submissions, suggesting that persuasive pleadings can channel both human and machine reasoning to similar ends. These findings support the potential of AI as a judicial aid but reinforce the need for robust human oversight.

Keywords: Artificial Intelligence, Judicial Drafting, AI Judgment, Legal drafting

1 Introduction

“If justice delayed is justice denied, then what of justice automated?”

In an era marked by digital acceleration and institutional strain, courts around the world are exploring how Artificial Intelligence (AI) might assist with the burdens of adjudication. Large language models (LLMs) like OpenAI's Generative Pre-trained Transformer (ChatGPT)¹—trained on a vast corpora of legal and general texts—are now widely used for research, drafting documents, and summarising tasks in legal

¹ See OpenAI “ChatGPT Overview” <https://openai.com/chatgpt/overview/> 2025 (last accessed 2025-03-24).

practice.² But can such a model go further? Can it draft a legal judgment of sufficient quality to meet the standards of a real court?

This article takes up that challenge by empirically evaluating the performance of ChatGPT-4o—the flagship large language model (LLM) from OpenAI—³on a real South African High Court case: *De Jager v Netcare Ltd.*⁴ While the scope of a single single case study is narrow, it does provide some useful initial insights into AI-driven judicial work future expanded experiments can build upon. The court case entailed constitutional privacy rights, statutory data protection law, and evidentiary admissibility. We asked ChatGPT-4o to write a full judicial opinion using real court filings from the case and a carefully constructed 20-step prompting protocol. We then compared its output to the actual judgment handed down by the presiding officer Motha J.

The stakes of this inquiry are high. In South Africa, judicial authority is constitutionally vested in the courts, and no person or organ of state may interfere with their functioning.⁵ Any suggestion that a machine might perform judicial reasoning raises profound constitutional, ethical, and institutional questions. Dismissing the potential of generative AI altogether would ignore the reality that such tools are already entering legal practice—including behind the bench.⁶

This article makes three contributions. First, it offers the first documented South African experiment in using an LLM to write a full-length legal judgment based on real case materials. Second, it develops a value-driven, replicable methodology for prompting and evaluating AI-generated judgments, grounded in local jurisprudence, judgment-writing pedagogy, and constitutional norms. Third, it provides a comparative analysis of the strengths and limitations of AI-generated judicial writing, identifying areas where such tools may support judicial work—particularly in drafting, accessibility, and structural clarity—and where they fall short, particularly in legal reasoning, citation, and contextual understanding.

Our central argument is that while ChatGPT-4o can generate judgments that are factually plausible and structurally coherent, it cannot yet replace human judicial reasoning. However, it holds significant promise as a judicial drafting assistant—a tool to accelerate workflow, support multilingual access, and enhance the transparency and traceability of the judgment-writing process. We conclude that South African courts should treat AI-generated outputs analogously to research clerks' memos: useful, suggestive, but never dispositive.

In the sections that follow, we review the legal and technological context, describe the experimental methodology, outline our evaluation criteria, and analyse the results.

² See Galarreta, Francisco Javier Fernández, "The Impact of Artificial Intelligence Technologies on the Justice Administration and on the Judicial Office Personnel" *Zeitschrift für Rechtssoziologie*, vol. 45, no. 1, 2025, pp. 103-121.

³ See OpenAI "ChatGPT-4o" <https://platform.openai.com/docs/models/gpt-4o> 2025 (last accessed 2025-03-13).

⁴ *De Jager v Netcare Limited and Others* 2025 JDR 0793 (GP).

⁵ S165 (1) of The Constitution of the Republic of South Africa, 1996.

⁶ Galarreta, Francisco Javier Fernández, "The Impact of Artificial Intelligence Technologies on the Justice Administration and on the Judicial Office Personnel" *Zeitschrift für Rechtssoziologie*, vol. 45, no. 1, 2025, pp. 103-121.

In doing so, we aim not to speculate, but to provide evidence-based guidance for how courts might pragmatically and constitutionally embrace AI as part of their evolving institutional toolkit.

2 Background and literature context

The integration of artificial intelligence (AI) into legal systems has shifted from theoretical curiosity to practical inevitability. Around the world, courts and legal practitioners are experimenting with large language models (LLMs) to streamline administrative tasks, conduct legal research, and in some instances, assist with decision-making. In Colombia, for example, a judge openly cited ChatGPT-3.5 in a judicial ruling, sparking international debate on transparency, accountability, and the limits of AI-generated reasoning.⁷ In South Africa, legal firms such as Schindlers have tested proprietary AI tools to simulate adjudication, and academic scholars have begun probing the normative boundaries of such use.⁸ Yet, the question remains: can generative AI ever satisfy the institutional and constitutional requirements of judicial authority?

The South African Constitution is unequivocal: judicial authority is vested in the courts and must be exercised independently and without fear, favour, or prejudice. Section 165(4) of the Constitution obliges the state to take “legislative and other measures” to support the independence and effectiveness of the judiciary. Whether generative AI may be considered such a “measure” raises urgent legal and ethical questions, particularly where automation could undermine impartiality, accountability, or the right to reasons. “To be democratically legitimate, rules and judgments should be the outcome of popular sovereignty, something over which the people exercise authorship.”⁹

Hicks et al. argue that generative LLMs are optimized for likelihood rather than likeness to reality and are trained to produce plausible-looking text, not to represent the world accurately so their errors (and even their correct answers) should be understood as “bullshit” outputs unconcerned with truth.¹⁰ This analogy underscores that accuracy is incidental under a truth-indifferent training regime because correctness is statistically emergent rather than epistemically grounded. This paper argues that assuming inaccuracy at the macro level may be a misstep if the goal is to guide responsible AI deployment. Such pessimism is, by definition, indifferent to AI as an assistive judicial tool—

⁷ See First Circuit Court of Cartagena, Case No. 11001-22-2023-00016.

⁸ See Schindlers Attorneys, ‘Sarah – Schindlers’ Adjudication AI Engine Hands Down Judgment in Favour of the Legalisation of Psilocybin (Psychedelic Mushrooms) in South Africa’ 22 April, <https://www.schindlers.co.za/sarah-schindlers-adjudication-ai-engine-hands-down-judgement-in-favour-of-the-legalisation-of-psilocybin-psychadelic-mushrooms-in-south-africa/> (last accessed 2025/05/15).

⁹ Kiel Brennan-Marquez and Stephen Henderson, *Artificial Intelligence and Role-Reversible Judgment*, 109 J. CRIM. L. & CRIMINOLOGY 137 (2019) p152.

¹⁰ Hicks, M.T., Humphries, J., & Slater, J. (2024). *ChatGPT is bullshit. Ethics and Information Technology*, 26, 38. <https://doi.org/10.1007/s10676-024-09775-5>.

one that would serve to support, rather than to make final binding decisions. While AI in law may on the surface appear epistemically hazardous, independent judicial authority requires that its outputs be validated by humans, just as one would evaluate the contributions of a research assistant.

Despite these high stakes, there is little empirical research in the South African context assessing how AI tools perform when applied to the task of judgment writing. Most local commentary to date has focused on AI in legal practice more broadly—including contract generation, predictive analytics, and document review. For instance Petse and Phindelo deemed ChatGPT a deceptive tool that threatens courts and the administration of justice mainly on the basis of three cases involving lawyers who submitted unverified AI-written submissions.¹¹ Their analysis treated a few instances as proof of systemic danger despite empirical studies proving that error-rates can be significantly reduced when users apply best practices¹²—such as supplying sources using Retrieval Augmented Generation (RAG).¹³ RAG entails prompting a large language model to generate an answer based only on supplied documents anchoring the response of the model to those documents.¹⁴ In the three cases, the courts were critical of the lawyers for copying ChatGPT in bad faith but did not condemn ChatGPT itself as an assistive tool.¹⁵ Thaldar also demonstrated that ChatGPT-4 can assist in drafting Data Transfer Agreements for health research but found its outputs uneven and in need of expert revision.¹⁶ Similarly, Lefakane explored the potential for AI to alleviate burdens on South Africa's overextended judiciary, concluding that any implementation must be directed by constitutional values and subject to rigorous oversight—but did not test AI performance on a real case.¹⁷

Internationally, researchers have experimented with LLMs by posing legal problems from a variety of jurisdictions. Broyde tested ChatGPT-4's ability to respond to complex halachic legal questions, finding that while the model could apply precedent and analyse texts, it struggled with contextual reasoning and ethical nuance.¹⁸ In a German study, Schweitzer and Conrads found that ChatGPT-4 could resolve basic business law

¹¹ I Petse & U Phindelo 'The Influence of ChatGPT-generated Data on the Administration of Justice in South Africa' (2025) 5 *Turf Law Journal* 1-14.

¹² B Tomlinson, AW Torrance & RW Black 'ChatGPT and Works Scholarly: Best Practices and Legal Pitfalls in Writing with AI' (2023) 76 *SMU L Rev Forum* 108.

¹³ See Reece Rogers 'Reduce AI Hallucinations With This Neat Software Trick' 2024 https://www.wired.com/story/reduce-ai-hallucinations-with-rag/?utm_source=chatgpt.com (last accessed 2025/06/12).

¹⁴ Ibid.

¹⁵ See *Mavundla v MEC: Department of Co-Operative Government and Traditional Affairs KwaZulu-Natal and Others* [2025] 2 ZAKZPHC para 37; *Parker v Forsyth* (2023) 1 1585/20 ZAGPRD para 89-90; *Mata v Avianca Inc* 22-cv-1461 (PKC) United States District Court, S.D. New York (2023).

¹⁶ See DW Thaldar 'How effectively can ChatGPT-4 draft data transfer agreements for health research?' (2025) *Humanit Soc Sci Commun* 12, 524.

¹⁷ N Lefakane 'The Use of Artificial Intelligence for Adjudication in South Africa' LLM dissertation, North-West University, 2022.

¹⁸ See Broyde & Michael J 'AI and Jewish Law: Seeing How ChatGPT 4.0 Looks at a Novel Issue' (2023) Buchman Law School, Tel Aviv University, and Cardozo Law School, Yeshiva University, *CSLR Research Paper No. 12.2023-AFF*.

disputes but lacked sophistication in edge cases.¹⁹ Collectively, these studies suggest that LLMs can simulate legal reasoning under constrained conditions but remain dependent on careful prompting, verification, and human correction.

African perspectives remain limited in this global discourse. While the 2025 Africa AI Privacy Report promotes AI use in transcribing rulings and managing caseloads, it does not substantively address whether AI can, or should, write judgments.²⁰ Nor does it engage with the jurisprudential traditions—such as ubuntu or restorative justice—that shape African legal reasoning and which may not be easily modelled by non-African AI systems trained primarily on Euro-American corpora.²¹

This article seeks to fill that gap. By subjecting ChatGPT-4o to a demanding, real-world test using South African case law, procedural rules, and constitutional values, we explore both its technical capabilities and its jurisprudential limitations. Our findings contribute to a grounded, context-sensitive understanding of AI's role in judicial writing—and offer practical insight for courts, academics, and policy-makers navigating the promises and perils of automation in the judiciary.

3 Methodology

This study employed an empirical case study approach to explore the capabilities of ChatGPT-4o, a large language model developed by OpenAI, in drafting a High Court judgment for a South African case, *De Jager v Netcare Ltd.*²² Selected for its doctrinal complexity and constitutional significance, the case concerned the admissibility of surveillance evidence in a delictual damages claim, raising issues under privacy provision Section 14 of the Constitution of the Republic of South Africa, 1996, and the Protection of Personal Information Act (POPIA).²³ By comparing the AI-generated judgment to the human-authored judgment by Motha J, the study assessed formal attributes—structure, factual accuracy, use of legal authority, legal reasoning, and clarity—as outlined in the evaluation framework (Section 4) below. The methodology prioritised transparency and alignment with South African judicial standards, focusing on the AI's ability to produce a plausible draft judgment while adhering to ethical and constitutional principles.

3.1 Case selection

The case, heard in the Gauteng Division of the High Court, Pretoria, involved a plaintiff claiming approximately R25.7 million in damages from a healthcare provider for negligent surgery. The central legal issue was whether the defendant's covert

¹⁹ S Schweitzer & M Conrads 'Correction to the digital transformation of jurisprudence: an evaluation of ChatGPT-4's applicability to solve cases in business law' (2024) *Artif Intell Law*.

²⁰ See Africa AI Policy Lab *Africa AI Privacy Report 2025* (2025) available at <https://www.aipolicy.africa/africa-intelligence-report> (last accessed 15 April 2025/04/15) 67, 73–74.

²¹ Ibid.

²² *De Jager v Netcare Limited and Others* 2025 JDR 0793 (GP).

²³ Act 4 of 2013.

surveillance—which incidentally captured the plaintiff’s minor grandchildren—violated privacy rights under Section 14 and POPIA, and whether the evidence was admissible. The case’s reliance on written submissions, including those from amici curiae (Professors Thaldar and Snail ka Mtuze), and its well-defined legal questions made it an ideal test case for AI judgment drafting. Two published judgments—a procedural ruling (2024)²⁴ and the substantive judgment (2025)²⁵—provided the basis for comparison, supporting the study’s exploratory aim to evaluate the AI’s formal drafting ability against a human judgment.

3.2 Large language model: ChatGPT-4o

Released in May 2024,²⁶ ChatGPT-4o was selected for its advanced multimodal capabilities and strong contextual coherence, despite lacking dedicated training in South African law.²⁷ ChatGPT-4o’s knowledge cut-off date is October 2023.²⁸ The experiment was conducted in a closed system using the RAG approach. The model relied exclusively on the provided case materials and prompts to generate its output whilst restricted from internet access. This simulated realistic judicial drafting conditions and minimised external influence. Newer models have since been released i.e. GPT-5. Methodology applied in this case study may not necessarily be replicable.

3.3 Prompting protocol

A structured 20-prompt and responses sequence (fully available in citation) was used to guide ChatGPT-4o in drafting the judgment.²⁹ This sequence emulated a High Court judge’s process of synthesising legal submissions to ensure effectiveness in accordance with effective prompting practice.³⁰ While the sequence included preparatory prompts (e.g., for document analysis), the central prompt instructed the model to produce a full-length judgment, using explicit parameters to ensure adherence to judicial standards in South Africa (discussed under subheading 4. Formulating evaluation criteria). The first and main prompt, read as follows:

“Prompt 1: Hello, ChatGPT. In the case of De Jager v Netcare, before the Pretoria High Court, a constitutional issue has been raised by the plaintiff concerning the surveillance of the plaintiff and some of his family members by private investigators appointed by the defendant. I would like you to draft a judgment on this constitutional

²⁴ *De Jager v Netcare* (2024) Case No: 42041/16 ZAGPPHC 503.

²⁵ *De Jager v Netcare Limited and Others* 2025 JDR 0793 (GP).

²⁶ See Kyle Wiggers ‘OpenAI debuts GPT-4o ‘omni’ model now powering ChatGPT’ *TechCrunch* 13 May 2024 <https://techcrunch.com/2024/05/13/openais-newest-model-is-gpt-4o/> (last accessed 12 June 2025/06/12).

²⁷ See OpenAI ‘What is the ChatGPT model selector?’ https://help.openai.com/en/articles/7864572-what-is-the-chatgpt-model-selector#h_80fc86948a (last accessed 2025/06/12).

²⁸ OpenAI ‘ChatGPT-4o’ <https://platform.openai.com/docs/models/gpt-4o> (last accessed 2025/03/13).

1. ²⁹ AI Judgment <https://chatgpt.com/share/68e29955-19ac-800d-9c63-cb0fd9f636cc> (last accessed 2025/10/5).

³⁰ See Z Lin ‘How to write effective prompts for large language models’ (2024) 8 *Nature Human Behaviour* 614.

issue. I will provide you with all the relevant court papers which contain written arguments and the case law cited in those arguments. Your role will be to act as the presiding judge, Judge Motha of the Pretoria High Court. You are to draft a judgment in the typical structure (provide heading & numbered paragraphs) and style of South African jurisprudence as reflected in the Juta Judgment Style Guide which I have provided you in the attachment. You must provide full citations of law & cases you utilise using footnotes; if you cannot perform this function provide references in brackets within the text. This task will entail, most prominently, that you provide an overview of the facts, analyse the issue for determination—i.e., the constitutional issue—critically evaluate the arguments presented by the parties and the *amicus curiae*, and decide the issue, with an appropriate cost order. You are required to make decisions based on papers provided and not your inherent knowledge. In writing your judgement you must consider judgement writing guidelines under South African law that are listed below: Write using a logical structure Provide reasons to justify your conclusions Be impartial and bias free Avoid undue technicality but maintain professionalism Utilise prescribed formats Succinctly deliberate and pronouns on all arguments, evidence, applicable laws Deliver judgement promptly Promote human rights subject to appropriate limitations We will now proceed as follows: I will provide you with ten of the relevant documents, one at a time. For each document, you will first provide me with a summary. Are you ready to begin?³¹

3.4 Inputs and exclusions

The AI was provided with the following documents: plaintiff's heads of argument, defendant's heads of argument, both parties' supplementary heads of argument, the *amicus curiae* submissions from Professors Thaldar and Snail ka Mtuze, the Rule 16A notice,³² relevant extracts from the Uniform Rules of Court (Rules 16A, 36, and provisions on admissibility),³³ and the *Juta Judgment Style Guide*.³⁴ These materials encompassed the legal and factual framework of the case. Excluded materials included oral submissions and transcripts, which were unavailable and less central to the study's aim of evaluating formal drafting. To prevent bias and regurgitation, the human-authored judgment was withheld during drafting.

³¹ AI Judgment <https://chatgpt.com/share/68e29955-19ac-800d-9c63-cb0fd9f636cc> (last accessed 2025/10/5).

³² Plaintiff's Notice in Terms of Rule 16A (24 May 2024).

³³ Supreme Court Act 59 of 1959 as amended. GNR 5124, G. 51056 of 16 August 2024 (with effect from 20 September 2024).

³⁴ See Juta 'Judgment Style Guide Juta' https://juta.co.za/media/filestore/2017/05/Judgment_Style_Guide_Juta.doc (last accessed 2025/01/29).

3.5 Ethical safeguards and constitutional alignment

To ensure constitutional alignment, the prompting protocol included explicit ethical directives. ChatGPT-4o was instructed to apply legal principles impartially,³⁵ justify conclusions logically, promoting constitutional values and rights (i.e. human dignity, equality, children's rights),³⁶ and acknowledge its analytical limitations. These directives were reiterated throughout the prompting sequence to align the AI's output with the expectations of S165(2) of the Constitution of the Republic of South Africa, 1996.³⁷ In embedding these safeguards, the study ensured that the AI remained an impartial support tool—capable of assisting and modelling after judicial reasoning norms such as effectiveness and accessibility.³⁸

4 Formulating Evaluation framework

To meaningfully assess ChatGPT-4o's performance in drafting a South African High Court judgment, this study developed a structured evaluation framework aligned with both doctrinal expectations and constitutional values. The framework draws on effective prompting mechanisms,³⁹ judgment-writing manuals,⁴⁰ ethical standards for legal professionals, and academic literature on AI-assisted legal reasoning. It comprises five interrelated criteria.

4.1 Structure

Judgments must be intelligible and adhere to recognised conventions. South African courts typically adopt a fluid format based on: introduction, factual background, issues for determination, legal framework, application of the law, and conclusion.⁴¹ Formal features such as paragraph numbering, proper citation, and meaningful headings contribute to navigability and professional integrity.⁴² The *Juta Judgment Style Guide*⁴³ and

³⁵ S165(2) of the Constitution of the Republic of South Africa, 1996: ‘The courts are independent and subject only to the Constitution and the law, which they must apply impartially and without fear, favour or prejudice’.

³⁶ S28(2) of the Constitution of the Republic of South Africa, 1996 ‘A child's best interests are of paramount importance in every matter concerning the child.’

³⁷ S28(2) of the Constitution of the Republic of South Africa, 1996 ‘A child's best interests are of paramount importance in every matter concerning the child.’

³⁸ Section 165(4) of the Constitution of the Republic of South Africa, 1996.

³⁹ Z Lin ‘How to write effective prompts for large language models’ (2024) 8 *Nature Human Behaviour* 611–615.

⁴⁰ Gauteng Local Division: Pretoria ‘Practice Manual of the Gauteng Division Pretoria’ 25 July 2011 <https://johannesburgbar.co.za/practice-manual> (last accessed 2025/03/16).

⁴¹ L Theron ‘Refreshing And Enhancing Judgment Writing Skills’ *The South African Judicial Education Journal* Volume 5 Issue 1 2022 Pages 4–5, Section IV.

⁴² B Tomlinson, AW Torrance & RW Black ‘ChatGPT and Works Scholarly: Best Practices and Legal Pitfalls in Writing with AI’ (2023) 76 *SMU L Rev Forum* 117

⁴³ Juta ‘Judgment Style Guide Juta’ https://juta.co.za/media/filestore/2017/05/Judgment_Style_Guide_Juta.doc (last accessed 2025/01/29).

the training materials of the South African Judicial Education Institute (SAJEI) were used as benchmarks.⁴⁴

- This criterion evaluates whether ChatGPT-4o:
- Organises content according to conventional judicial structure;
 - Applies paragraph numbering correctly;
 - Uses appropriate citation format;
 - Produces a document consistent with professional publication standards.

4.2 Factual accuracy

Judicial decisions must be grounded in a faithful representation of the factual matrix. Misstatements of fact, incorrect attribution of arguments, or distortion of the procedural history can compromise the integrity of a judgment.⁴⁵ In the context of generative AI, the risk of hallucinated or fabricated content must be carefully assessed and verified with independent reading of original sources.

- This criterion evaluates whether ChatGPT-4o:
- Accurately identifies and summarises material facts from the record;
 - Correctly attributes arguments and evidence to the appropriate party;
 - Avoids factual invention, distortion, or omission.

4.3 Use of legal authority

Value 6 of the Bangalore Principles places emphasis on making conscientious conclusions based on understanding of the law.⁴⁶ Thus citing relevant legislation, precedents, and scholarly sources is central to sound judgment writing.⁴⁷ Proper citation affirms adherence to the principle of legality, ensures transparency, reinforces the persuasive weight of judicial decisions and affirms faith in the integrity of the judicial decisions.⁴⁸ In this context, authority is assessed not only for correctness but also for contextual appropriateness.

- This criterion evaluates whether ChatGPT-4o:
- Correctly identifies and applies relevant statutory provisions (e.g. POPIA, Constitution);
 - Cites applicable judicial precedents from relevant South African courts;
 - Appropriately draws on legal texts and academic commentary;
 - Demonstrates an understanding of how and why legal authorities are invoked.

⁴⁴ Established in terms of South African Judicial Education Institute Act 14 of 2008.

⁴⁵ B Tomlinson, AW Torrance & RW Black ‘ChatGPT and Works Scholarly: Best Practices and Legal Pitfalls in Writing with AI’ (2023) 76 *SMU L Rev Forum* 108.

⁴⁶ United Nations Bangalore Principles and Commonwealth Latimer House Principles ECOSOC 2006/23.

⁴⁷ B Tomlinson, AW Torrance & RW Black ‘ChatGPT and Works Scholarly: Best Practices and Legal Pitfalls in Writing with AI’ (2023) 76 *SMU L Rev Forum* 117.

⁴⁸ United Nations Bangalore Principles and Commonwealth Latimer House Principles ECOSOC 2006/23 Value 6.

4.4 Legal reasoning

Legal reasoning constitutes the core of any judicial opinion and ‘...the primary purpose of the judgment is to communicate the decision as well as the reasoning of the court.’⁴⁹ Legal reasoning involves the interpretation and application of legal rules to the facts, the assessment of competing arguments, and the articulation of a logically coherent path to judgment.⁵⁰ The purpose here is not to determine whether the AI or the court reached the “correct” outcome, but to assess, in a formal sense, the internal consistency and logical progression of the reasoning presented.

This criterion evaluates whether ChatGPT-4o:

Clearly identifies and frames the legal questions in dispute;

Applies relevant legal principles and interpretive doctrines (e.g. subsidiarity, proportionality, evidentiary thresholds);

Engages with the core submissions of both parties and any amici curiae;

Demonstrates logical coherence, doctrinal fidelity, and analytical adequacy.

By isolating the structural quality of legal reasoning from its normative content, this criterion allows for an assessment of the model’s capacity to simulate formal judicial logic—not its capacity for judicial discretion or moral judgment.

4.5 Clarity

‘Brevity, simplicity, and clarity are the watchdogs for effective judicial writing.’⁵¹ Essentially judicial language must be clear, internally consistent, and appropriate to its legal and institutional function, courteous and averse to imploring inflammatory or disparaging language.⁵² Judgments are written primarily for legal professionals and future courts, and may appropriately employ legal terminology, including Latin maxims and technical phrases. The expectation is not simplicity but clarity—the avoidance of unnecessary convolution, ambiguity, or verbosity. This approach reflects the realities of South African legal writing, where technical language is expected and where clarity is judged not by popular accessibility, but by the internal communicative standards of the legal profession.

This criterion evaluates whether ChatGPT-4o:

⁴⁹ L Theron ‘Refreshing And Enhancing Judgment Writing Skills’ *The South African Judicial Education Journal* Volume 5 Issue 1 2022 Page 3, Section IV.

⁵⁰ L Theron ‘Refreshing And Enhancing Judgment Writing Skills’ *The South African Judicial Education Journal* Volume 5 Issue 1 2022 Pages 4–5, Section IV.

⁵¹ Theron citing address delivered by the former Chief Justice Ngcobo of the Constitutional Court of South Africa entitled ‘Judgment Writing’ at L Theron ‘Refreshing And Enhancing Judgment Writing Skills’ *The South African Judicial Education Journal* Volume 5 Issue 1 2022 9.

⁵² ‘Do be respectful and courteous to all parties, the legal representatives as well as colleagues’ at L Theron ‘Refreshing And Enhancing Judgment Writing Skills’ *The South African Judicial Education Journal* Volume 5 Issue 1 2022, 8.

Uses language that is appropriate to the judicial context;
Presents its reasoning in a coherent and comprehensible manner;
Maintains a tone of judicial professionalism and impartiality;
Avoids stylistic inconsistency or rhetorical excess.

5 Results and comparative analysis

The 20-prompt experiment in ChatGPT-4o generated a full-length judgment (AI Judgment) in the Appendix—which was assessed against the actual judgment in *De Jager Case* by Motha J. Applying the evaluation criteria established in Section 4, the comparison confirmed the AI-generated judgment's superior technical performance across four core areas—structure, factual accuracy, legal reasoning, and clarity—while recognising the human-authored judgment's effective use of expert input and its engagement with broader contextual considerations.

5.1 Structure

The AI-generated judgment broadly followed the conventional structure of a South African High Court judgment, presenting distinct sections titled Introduction, Factual Background, Legal Framework, Arguments Presented, Evaluation of Arguments, and Conclusion. It employed consistent paragraph numbering, clearly defined headings, and an overall tone aligned with the Juta Judgment Style Guide. Case and statutory citations—such as *Bernstein v Bester*⁵³ and relevant POPIA provisions—were accurately cited. The inclusion of a formal order and a judge's signature block further contributed to its professional appearance. However, certain stylistic choices deviated from South African judicial norms. The use of bullet points (e.g., to enumerate legal questions and to structure the order), bold text for headings (e.g., “Introduction,”), and emphasis (e.g., “reasonable expectation of privacy,”) are inconsistent with established High Court conventions, which favour numbered paragraphs and plain or underlined headings.⁵⁴ While these elements may enhance readability in general prose, they are inappropriate for formal judgments and detract from stylistic fidelity. Additionally, minor repetition of central issues—such as privacy and admissibility—slightly affected navigability.

The human-authored judgment presented a broadly recognisable structure, with sections titled Introduction, The Issue in Contention, Submissions by Amici Curiae and Counsel, Under POPIA, Conclusion, and Order.⁵⁵ It did not include a dedicated factual background section, and legal analysis was often blended with the legal framework, which reduced clarity. Transitions between key issues were occasionally abrupt or implicit. It correctly employed paragraph numbering, footnoting and conformed to basic

⁵³ 1996 (2) SA 751 (CC).

⁵⁴ *AI Judgment*, para 24; para 76; para 1; para 50.

⁵⁵ *De Jager v Netcare Ltd* para 1; para 5; para 11; para 20; para 31; para 32.

formatting expectations in line with SAJEI standards. Importantly, it avoided the use of bullets or bold text, adhering more closely to conventional judicial presentation.

Comparative Assessment: The AI judgment exhibited strong internal structure and logical sequencing, but its use of non-standard formatting elements—such as bullets and boldface—undermined its conformity to High Court conventions. The Human Judgment, though less well-structured and occasionally lacking in clarity, adhered more faithfully to accepted formatting norms. As a result, it holds a slight advantage in stylistic compliance, despite its structural limitations.

5.2 Factual accuracy

The AI-generated judgment presented a detailed and factually accurate account of the background relevant to the constitutional challenge. It correctly identified the surveillance conducted by Dion Pienaar, the plaintiff's R25.7 million damages claim, and the incidental inclusion of minor grandchildren in the surveillance footage.⁵⁶ The judgment appropriately attributed the plaintiff's reliance on Section 14 of the Constitution to support a claim of unlawful surveillance, and the defendant's counter-argument that the surveillance was justified for litigation purposes.⁵⁷ The procedural history was clearly and coherently summarised.⁵⁸ Although the judgment included extraneous factual details—such as the plaintiff's cataract surgery and the specific timeline of the surveillance (August–September 2020)—these were accurate, if not strictly necessary for resolving the constitutional and statutory questions.⁵⁹

The human-authored judgment also accurately referenced the central facts necessary for adjudicating the constitutional and statutory issues.⁶⁰ It noted the surveillance conducted by Dion Pienaar and the plaintiff's damages claim, expressed within a range of R24.8–25.7 million.⁶¹ While it did not contain a dedicated factual section or elaborate on the broader factual context (e.g., the surgery or precise surveillance timeline), these omissions were not material to the legal reasoning, which focused principally on the application of subsidiarity and the interpretation of POPIA. The limited engagement with party submissions reflects the procedural posture of the case: the defendant adopted the submissions of Professor Thaldar without elaboration, and the plaintiff only addressed POPIA in supplementary heads of argument, prompted by Thaldar's intervention. The judgment's reliance on *amici curiae*—Professors Thaldar and Snail ka Mtuze—provided the necessary interpretive structure, particularly in light of the parties' constrained submissions. No factual errors or misrepresentations were identified.

Comparative Assessment: Both judgments satisfied the accuracy criterion, presenting reliable and contextually appropriate factual accounts. The AI Judgment held a slight advantage due to its clearer attribution of arguments and fuller factual detail,

⁵⁶ *AI Judgment* para 7.

⁵⁷ *AI Judgment* para 14–16; para 53; para 75.

⁵⁸ *AI Judgment* para 1–12.

⁵⁹ *AI Judgment* para 6–8.

⁶⁰ *De Jager v Netcare Ltd* para 5.

⁶¹ *De Jager v Netcare Ltd* para 36.

while the Human Judgment offered sufficient but more concise references, appropriately reflecting its doctrinal focus and reliance on amici for interpretive clarity.

5.3 Use of legal authority

The AI judgment cited an array of legal authorities, including constitutional provisions (sections 14, 28, and 36), various sections of POPIA (e.g., 11, 12, 18, 27, 34), and leading case law such as *Bernstein v Bester*,⁶² *Gaertner v Minister of Finance*,⁶³ *Key v Attorney-General*,⁶⁴ and *City of Cape Town v SANRAL*.⁶⁵ These citations were generally contextually appropriate and well integrated into the judgment's logical flow. For example, *Bernstein* was used to demarcate the scope of privacy rights, particularly for surveillance in private settings, while *Gaertner* supported the assessment of reasonable expectations of privacy in the home.⁶⁶ *Key* framed the admissibility of potentially unlawfully obtained evidence, and *City of Cape Town* supported the balance between privacy and open justice.⁶⁷ The AI judgment also referenced *Eskom Holdings SOC Ltd v Vaal River Development Association (Pty) Ltd and Others* to support the application of the subsidiarity doctrine.⁶⁸ The authorities were formatted correctly and used with superficial coherence. However, the judgment did not cite any academic literature, limiting its depth of engagement with broader theoretical discourse.

The human-authored judgment cited a wider selection of legal authorities. It relied on constitutional provisions (sections 14 and 36), POPIA sections (including 2, 6, 9, 11, 18, 26, 27), and precedent such as *Bernstein v Bester*, *Mazibuko v City of Johannesburg*,⁶⁹ *Eskom Holdings*, *Gaertner*, and *Smuts⁷⁰ v Botha*.⁷¹ The use of *Mazibuko* and *Eskom Holdings* appropriately grounded the subsidiarity argument, emphasising that POPIA, rather than a direct constitutional claim, was the proper basis for the plaintiff's challenge.⁷² *Smuts v Botha* was cited to assess the reasonableness of public surveillance, aligning with the case's focus.⁷³ Academic works by Cohen and Neethling, enhanced the judgment's doctrinal context, particularly in the analysis of personality rights.⁷⁴ Judge Motha referenced the European General Data Protection Regulation 2018 (GDPR) and traces its lineage from German models.⁷⁵ The amici

⁶² (1996 (2) SA 751 (CC);
AI Judgment para 26, 39, 50.

⁶³ 2014 (1) SA 442 (CC).

⁶⁴ 1996 (4) SA 187 (CC).

⁶⁵ 2015 (3) SA 386 (SCA).

⁶⁶ *AI Judgment* 15–16.

⁶⁷ *AI Judgment* para 47; 65.

⁶⁸ 2023 (4) SA 325 (CC).

⁶⁹ 2010 (4) SA 1 (CC).

⁷⁰ 2022 (2) SA 425 (SCA).

⁷¹ *De Jager v Netcare Ltd* para 2–3; para 6–8; para 11, para 14–15, 17–19; para 26–27; para 34.

⁷² *De Jager v Netcare Ltd* para 5–8.

⁷³ *De Jager v Netcare Ltd* para 34.

⁷⁴ *De Jager v Netcare Ltd* para 2; para 31.

⁷⁵ *De Jager v Netcare Ltd* para 38.

curiae—Professors Thaldar and Snail ka Mtuze—provided significant interpretive contributions, particularly in clarifying the applicability of POPIA exemptions and reinforcing the importance of subsidiarity. Their input was procedurally significant, as the defendant adopted Thaldar’s submissions and the plaintiff addressed POPIA only belatedly. However, the judgment did not reference certain evidentiary precedents such as *S v Naidoo*, which might have strengthened the admissibility analysis. The judgment also included a brief discussion on decolonisation and Afro-centric law that, while socially relevant, lacked legal specificity and did not contribute to doctrinal analysis.⁷⁶

Comparative Assessment: The AI judgment demonstrated imprecise and informal citation of relevant legal authorities, offering structurally sound integration of constitutional principles and precedent. Its omission of academic commentary limited its interpretive richness. The Human Judgment, while citing fewer cases overall, enriched its analysis through scholarly references and the strategic and procedurally justified use of amici submissions. Its citation of *Smuts v Botha* was appropriate and well-contextualised. However, the absence of additional evidentiary case law and the inclusion of a rhetorically framed decolonisation discussion somewhat diluted the legal focus. On balance, the AI judgment showed greater formal breadth in its citation practices, while the Human Judgment demonstrated complementary intellectual and procedural value. This evaluation is limited to the formal application of legal authorities and does not assess their substantive correctness.

5.4 Legal reasoning

The AI judgment clearly articulated the legal questions: whether the surveillance infringed the plaintiff’s constitutional right to privacy under section 14, whether it complied with POPIA, and whether the evidence was admissible.⁷⁷ It demonstrated a formal understanding of subsidiarity, citing *Eskom Holdings SOC Ltd* to support the conclusion that POPIA, not the Constitution, should have been the primary basis for the plaintiff’s claim.⁷⁸ Its proportionality analysis under section 36 was logically structured, evaluating the necessity and reasonableness of the surveillance.⁷⁹ The AI judgment concluded that surveillance in public spaces was permissible due to a diminished expectation of privacy, whereas surveillance at the plaintiff’s home and involving minors was not, noting less intrusive alternatives like Rule 36 medical examinations.⁸⁰ It engaged with the parties’ submissions and incorporated amici perspectives, citing relevant precedent including *Bernstein v Bester, Gaertner*, and *Key v Attorney-General*, showing a superficially coherent application of legal principles. Minor repetition did not detract significantly from its analytical structure.

⁷⁶ *De Jager v Netcare Ltd* para 29; para 30.

⁷⁷ *AI Judgment* para 74.

⁷⁸ *AI Judgment* para 75.

⁷⁹ *AI Judgment* para 75.

⁸⁰ *AI Judgment* para 10; para 33; para 69; para 75.

The human-authored judgment focused on whether the plaintiff's section 14 claim was sustainable in light of the subsidiarity principle.⁸¹ Citing *Mazibuko*⁸² v *City of Johannesburg* and *Eskom Holdings*, it concluded that POPIA should have been relied on instead of the Constitution—a procedurally sound finding given that POPIA was only introduced via Professor Thaldar's amicus submission.⁸³ The defendant adopted these submissions without elaboration, and the plaintiff responded only in supplementary heads. The judgment's proportionality analysis under section 36 was brief but formally correct, referencing *Gaertner and Smuts v Botha*⁸⁴ to assess the legitimacy of the surveillance.⁸⁵ The reliance on amici curiae (Professors Thaldar and Snail ka Mtuze) was contextually appropriate and provided necessary doctrinal clarity. The conclusion to admit the surveillance under POPIA's litigation exemption was supported by a logical rationale. A discussion of decolonisation and Afro-centric legal theory, though present, was tangential and did not substantively contribute to the legal reasoning. Transitions between issues were less clearly structured, and the engagement with party submissions was minimal, but consistent with the case's procedural dynamics.

Comparative Assessment: The AI judgment exhibited greater structural coherence, clearer issue framing, and broader engagement with submissions and precedent, providing an advantage in formal reasoning. It demonstrated a superficially coherent application of key doctrines and maintained analytical adequacy throughout. The Human Judgment, while less systematically developed, correctly applied subsidiarity and POPIA principles and effectively relied on amici to address procedural deficiencies. However, its transitions were less clear, and the inclusion of unrelated socio-legal commentary somewhat diluted its analytical focus. On balance, the AI judgment reflected greater technical rigour, while the Human Judgment demonstrated sufficient doctrinal understanding within a narrower scope. This assessment is limited to the formal application of legal principles, without commenting on the substantive merits of either judgment's outcome.

5.5 Clarity

The AI judgment maintained a formal, professional, and impartial tone throughout. It used legal terminology with precision—for example, 'reasonable expectation of privacy', 'fruit of the poisonous tree', and 'proportionality under section 36'—in correct doctrinal contexts.⁸⁶ Complex constitutional and statutory issues were clearly articulated, with logical transitions between sections. Sentences were concise and unambiguous, and the final order was phrased with clarity and formality. The AI judgment's language use aligned well with formal judicial communication and would be readily intelligible to legal professionals.

⁸¹ *De Jager v Netcare Ltd* para 6–7.

⁸² 2010 (4) SA 1 (CC).

⁸³ *De Jager v Netcare Ltd* para 5–9.

⁸⁴ 2022 (2) SA 425 (SCA).

⁸⁵ *De Jager v Netcare Ltd* para 19–24.

⁸⁶ *AI Judgment* para 10–11; para 19; para 11.

The human-authored judgment communicated key legal issues effectively and used doctrinal language accurately. It correctly employed terms like ‘subsidiarity’, ‘*boni mores*’, ‘data subject’, and ‘minimality principle’ within their relevant contexts.⁸⁷ Key findings, particularly those addressing amici contributions and the final order, were expressed with clarity. However, the judgment’s writing style was uneven at times. Informal expressions such as “just what the doctor ordered,” “packs a devastating punch,” and “Kasi-flavour” disrupted the otherwise judicial tone.⁸⁸ The discussion of decolonisation and Afro-centric law had potential relevance to South Africa’s socio-legal context, but it was not connected to the legal analysis of privacy, POPIA, or admissibility.⁸⁹ As a result, it came across as rhetorical rather than substantive, diminishing cohesion. Some transitions between sections were less fluid, occasionally requiring the reader to infer logical connections. Despite these stylistic issues, the judgment successfully conveyed its core reasoning and conclusions.

Comparative Assessment: The AI judgment exhibited consistent formality, simpler legal terminology, and a professional tone suited to High Court writing. The human judgment, while generally clear in its analysis, suffered from uneven stylistic execution and rhetorical asides that detracted from coherence. Although the decolonisation commentary could have added socio-legal insight, its lack of actionable development rendered it a distraction. On balance, the AI judgment better conformed to the communicative expectations of judicial writing, while the human-authored judgment—though effective in parts—was less consistent in tone and structure. This assessment concerns the clarity and suitability of language, without evaluating the substantive merits of either judgment.

5.6 Overall assessment

This evaluation, focuses on formal logic and conformity to judicial standards, does not assess the substantive correctness of either outcome. Across the five evaluative criteria, the AI-generated judgment outperformed the human-authored judgment in structural coherence, factual accuracy, formal legal reasoning, and linguistic clarity, demonstrating strong technical alignment with the standards expected of a South African High Court judgment. Its structured layout, accurate use of legal authorities—such as *Bernstein v Bester* and *Key v Attorney-General*—and clear articulation of doctrines like subsidiarity and proportionality underscored its potential as an effective drafting tool. However, its use of non-standard formatting elements—namely bullets and bold text (e.g., headings)—departed from accepted judicial publishing norms, slightly diminishing stylistic fidelity. Occasional repetition also detracted from overall concision.⁹⁰

The human-authored judgment, while less technically refined, effectively leveraged the input of amici curiae—Professors Thaldar and Snail ka Mtuze—to address procedural shortcomings, particularly in light of the plaintiff’s delayed engagement with

⁸⁷ *De Jager v Netcare Ltd* para 7; para 33; para 35; para 34.

⁸⁸ *De Jager* para 1; para 5; para 38.

⁸⁹ *De Jager v Netcare Ltd* para 37.

⁹⁰ *AI Judgment* para 24; para 76; para 50.

POPIA and the defendant's adoption of Thaldar's submissions. Its citation of academic sources (e.g., Cohen, Neethling) and contextually appropriate case law (e.g. *Smuts v Botha*) added doctrinal depth. Nonetheless, structural inconsistencies, stylistic variability, and an underdeveloped decolonisation discussion reduced its overall formal coherence. While the commentary on decolonisation was potentially relevant to broader socio-legal concerns—such as colonial influences on privacy law—it was not substantively integrated into the legal analysis and appeared more rhetorical than doctrinal.

These findings highlight the significant promise of generative AI in producing structured, legally plausible draft judgments, particularly in technical execution. However, formatting irregularities illustrate the need for closer adherence to judicial conventions. The human judgment's strategic use of expert contributions and engagement with South Africa's procedural realities underscores the irreplaceable interpretive role of human judges in constitutional adjudication. AI-generated drafts may serve as valuable foundations for effective judicial writing, but rigorous human oversight remains essential to ensure stylistic compliance, doctrinal soundness and that judicial authority and independence remains with the courts in alignment with constitutional values, as required by section 165 of the Constitution.

6 Discussion

This study represents an initial exploration of generative AI, specifically the model ChatGPT-4o, in the context of judicial drafting, comparing its output to a human-authored High Court judgment in *De Jager v Netcare Ltd*. The evaluation focused on formal criteria—structure, factual accuracy, use of legal authority, legal reasoning, and clarity of language—and found that the AI-generated judgment outperformed the human-authored judgment in technical execution. Notably, however, both judgments exhibited striking analytical and conclusory similarities. This shared reasoning trajectory—driven by a common reliance on party and amici submissions—raises important questions about the role of legal argumentation in shaping both AI and human decision-making. This section explores the significance of these findings, their implications for legal technology, and directions for future research, while emphasising the exploratory nature of this inquiry.

6.1 Surprising convergence: Analytical and conclusory similarity

The most notable finding was the similarity between the AI and Human Judgments in both analytical structure and legal conclusions. Both judgments framed the key legal questions—constitutional privacy (Section 14), POPIA compliance, and evidence admissibility—through the doctrines of subsidiarity (*Eskom Holdings SOC Ltd; Mazibuko v City of Johannesburg*) and proportionality (Section 36), ultimately finding that surveillance in public spaces was admissible while surveillance of the home or minors was more problematic. Although their orders diverged slightly—the AI judgment explicitly excluded home/minor footage, while the Human Judgment admitted surveillance more

broadly under POPIA exemptions—both prioritised litigation imperatives over privacy claims.

This convergence is especially significant given the evaluation's methodological focus on formal qualities rather than substantive outcomes. As detailed in the *Results*, both judgments were shaped by the same argumentative inputs: the plaintiff's delayed invocation of POPIA, the defendant's adoption of Professor Thaldar's arguments, and the amici submissions of Professors Thaldar and Snail ka Mtuze. The *Legal reasoning* analysis further showed that both judgments drew from a shared corpus of precedents, such as *Bernstein v Bester* and *Gaertner v Minister of Finance*, reinforcing the analytical symmetry.

6.2 Why the Similarity? Submissions as a framing force

The similarity in reasoning likely reflects the central role of submissions in framing the analysis. As noted in *Factual accuracy*, the plaintiff focused initially on constitutional rights, while amici supplied the doctrinal framework of POPIA exemptions and subsidiarity. The defendant adopted Thaldar's submissions in full. The AI judgment systematically presented these arguments, while the Human Judgment relied on amici to clarify the legal framework.⁹¹

This convergence invites further reflection: do AI and human judges process persuasive legal arguments in comparable ways? In South African legal practice, well-developed submissions often shape outcomes. The AI's capacity to mirror this process—producing a formally coherent judgment grounded in the same inputs—suggests that AI may emulate core patterns of human legal reasoning when prompted with structured material. Alternatively, it may reflect a deeper alignment between the way large language models and judges weigh legal inputs. While these findings are preliminary, they raise meaningful questions about the nature of legal reasoning across human and artificial decision-makers.

6.3 Technical strengths and stylistic flaws of AI

The *Results* confirmed the AI judgment's technical strengths, including superior structural coherence (Section 5.1), detailed factual precision (Section 5.2), broad use of authority (Section 5.3), clear legal reasoning (Section 5.4), and consistent tone and clarity (Section 5.5). These traits mark generative AI as a promising tool for producing polished, legally plausible draft judgments. However, formatting deviations—particularly the use of bullets and bold text—departed from the *Juta Judgment Style Guide* and detracted from stylistic fidelity.⁹²

Additionally, the AI's lack of scholarly engagement (Section 5.3) highlights a limitation in interpretive richness. While its use of terminology such as "reasonable expectation of privacy" was contextually accurate, it offered limited insight into normative

⁹¹ *De Jager v Netcare Ltd* para 11–13.

⁹² *AI Judgment* para 3; para 8; para 10–13; para 17; para 21; para 33; para 42; para 75.

or academic debates.⁹³ This suggests that AI excels in formal replication of legal language but still requires human input to achieve interpretive depth.

6.4 Human Judgment's procedural and contextual nuances

Despite weaker technical execution, the Human Judgment demonstrated contextually informed decision-making. Its reliance on amici was appropriate, given the plaintiff's delayed POPIA submissions and the defendant's reliance on Thaldar. It effectively integrated academic commentary (e.g., Neethling, Cohen) and cited *Smuts v Botha* to assess surveillance reasonableness (Section 5.3).⁹⁴

However, structural inconsistency (Section 5.1), variable tone (Section 5), and the inclusion of rhetorical devices (e.g., "Kasi-flavour") reduced formal coherence.⁹⁵ The decolonisation discussion, while potentially relevant to critiques of colonial legacies in privacy law, was not tied to the legal analysis and functioned more as a rhetorical aside than a substantive argument (Section 5). This illustrates the human capacity to introduce socio-legal commentary, even when imperfectly executed—something AI cannot replicate without explicit prompting.

6.5 Implications and future directions

The analytical alignment between AI and human judgments in this case suggests that AI can draft judgments that are, at least formally, comparable to human output—particularly when prompted with high-quality submissions and quality prompts. While AI's formatting errors and shallow normative engagement remain significant limitations, the findings support the notion of AI as a valuable drafting tool.

Future studies should explore how generative AI performs across different legal contexts, including other jurisdictions, levels of court, and types of legal disputes. Comparative research on large language models and prompt engineering may further clarify whether AI genuinely mirrors human reasoning or merely assembles plausible text. The Human Judgment's use of amici also suggests a future role for AI in organising and structuring expert input for judicial review.

7 Conclusion

This study demonstrates that ChatGPT-4o can generate technically proficient, legally plausible judicial drafts. The AI judgment displayed strong formal structure, doctrinal accuracy, and clarity, though it was weakened by formatting inconsistencies and limited contextual depth. Strikingly, its analysis and conclusions closely mirrored those of the Human Judgment—likely a result of shared reliance on party and amici submissions. The Human Judgment, for its part, illustrated the enduring value of human discretion,

⁹³ *AI Judgment* para 10–11.

⁹⁴ *De Jager v Netcare Ltd* para 2; para 34.

⁹⁵ *De Jager* para 1; para 5; para 38.

expert engagement, and contextual awareness, despite structural and stylistic limitations. These findings support the potential of AI as a judicial aid but reinforce the need for robust human oversight to ensure doctrinal integrity and constitutional fidelity, as required by section 165(4) of the Constitution. The evaluation focuses on form, not substantive correctness; whether both judgments' conclusions were legally sound remains a question for appellate review.

8 References

1. Africa AI Policy Lab Africa AI Privacy Report 2025 (2025) available at <https://www.aipolicy.africa/africa-intelligence-report> (last accessed 2025/04/15).
2. Bernstein v Bester 1996 (2) SA 751 (CC).
3. Broyde, M & Michael, J 'AI and Jewish Law: Seeing How ChatGPT 4.0 Looks at a Novel Issue' (2023) Buchman Law School, Tel Aviv University, and Cardozo Law School, Yeshiva University, CSLR Research Paper No. 12.2023-AFF.
4. Hicks, M.T., Humphries, J., & Slater, J. (2024). *ChatGPT is bullshit. Ethics and Information Technology*, 26, 38. <https://doi.org/10.1007/s10676-024-09775-5>.
5. City of Cape Town v SANRAL 2015 (3) SA 386 (SCA).
6. De Jager v Netcare (2024) Case No: 42041/16 ZAGPPHC 503.
7. De Jager v Netcare Limited and Others 2025 JDR 0793 (GP)
8. Eskom Holdings SOC Ltd v Vaal River Development Association (Pty) Ltd and Others 2023 (4) SA 325 (CC).
9. European General Data Protection Regulation 2018 (GDPR).
10. First Circuit Court of Cartagena, Case No. 11001-22-2023-00016.
11. AI Judgment <https://chatgpt.com/share/68e29955-19ac-800d-9c63-cb0fd9f636cc> (last accessed 2025/10/5).
12. Gaertner v Minister of Finance 2014 (1) SA 442 (CC).
13. Galarreta, Francisco Javier Fernández, "The Impact of Artificial Intelligence Technologies on the Justice Administration and on the Judicial Office Personnel" *Zeitschrift für Rechtssoziologie*, vol. 45, no. 1, 2025.
14. Gauteng Local Division: Pretoria 'Practice Manual of the Gauteng Division Pretoria' 25 July 2011 <https://johannesburgbar.co.za/practice-manual> (last accessed 2025/03/16).
15. Hicks, M.T., Humphries, J. & Slater, J. ChatGPT is bullshit. *Ethics Inf Technol* 26, 38 (2024). <https://doi.org/10.1007/s10676-024-09775-5>
16. Juta 'Judgment Style Guide Juta' https://juta.co.za/media/filestore/2017/05/Judgment_Style_Guide_Juta.doc (last accessed 2025/01/29).
17. Key v Attorney-General 1996 (4) SA 187 (CC).
18. Kyle Wiggers 'OpenAI debuts GPT-4o 'omni' model now powering ChatGPT' TechCrunch 13 May 2024 <https://techcrunch.com/2024/05/13/openais-newest-model-is-gpt-4o/> (last accessed 2025/06/12)
19. KwaZulu-Natal and Others [2025] 2 ZAKZPHC.
20. Lin Z 'How to write effective prompts for large language models'(2024) 8 Nature Human Behaviour 614.
21. Mata v Avianca Inc 22-cv-1461 (PKC) United States District Court, S.D. New York (2023).
22. Mavundla v MEC: Department of Co-Operative Government and Traditional Affairs [2025] 2 ZAKZPHC.
23. Mazibuko v City of Johannesburg 2010 (4) SA 1 (CC)

24. N Lefakane ‘The Use of Artificial Intelligence for Adjudication in South Africa’ LLM dissertation, North-West University, 2022.
25. Open AI ‘ChatGPT Traceability Document - AI Experiment’ <https://chatgpt.com/canvas/shared/67cd52d06db08191a3507752fc022e1> (last accessed 2025/04/19).
26. OpenAI “ChatGPT Overview” <https://openai.com/chatgpt/overview/> 2025 (last accessed 2025-03-24).
27. OpenAI “ChatGPT-4o” <https://platform.openai.com/docs/models/gpt-4o> 2025 (last accessed 2025-03-13).
28. OpenAI ‘ChatGPT-4o’ <https://platform.openai.com/docs/models/gpt-4o> (last accessed 2025/03/13).
29. OpenAI ‘What is the ChatGPT model selector?’ https://help.openai.com/en/articles/7864572-what-is-the-chatgpt-model-selector#h_80fc86948a (last accessed 2025/06/12).
30. Parker v Forsyth (2023) 1 1585/20 ZAGPRD.
31. Petse, I & Phindelo, U ‘The Influence of ChatGPT-generated Data on the Administration of Justice in South Africa’ (2025) 5 Turf Law Journal 1-14.
32. Protection of Personal Information Act 4 of 2013.
33. Reece Rogers ‘Reduce AI Hallucinations With This Neat Software Trick’ 2024 https://www.wired.com/story/reduce-ai-hallucinations-with-rag/?utm_source=chatgpt.com (last accessed 2025/06/12).
34. S v Naidoo 1998 (1) SACR 479 (NPD)
35. Schindlers Attorneys, ‘Sarah – Schindlers’ Adjudication AI Engine Hands Down Judgment in Favour of the Legalisation of Psilocybin (Psychedelic Mushrooms) in South Africa’ 22 April 2024 <https://www.schindlers.co.za/sarah-schindlers-adjudication-ai-engine-hands-down-judgement-in-favour-of-the-legalisation-of-psilocybin-psychadelic-mushrooms-in-south-africa/> (last accessed 2025/05/15)
36. Schweitzer, S & Conrads, M ‘Correction to the digital transformation of jurisprudence: an evaluation of ChatGPT-4’s applicability to solve cases in business law’ (2024) Artif Intell Law.
37. Smuts v Botha 2022 (2) SA 425 (SCA).
38. Supreme Court Act 59 of 1959 as amended. GNR 5124, G. 51056 of 16 August 2024 (with effect from 20 September 2024).
39. Thaldar, DW ‘How effectively can ChatGPT-4 draft data transfer agreements for health research?’ (2025) Humanit Soc Sci Commun 12, 524.
40. The Constitution of the Republic of South Africa, 1996.
41. The South African Judicial Education Institute Act 14 of 2008.
42. Theron, L ‘Refreshing And Enhancing Judgment Writing Skills’ The South African Judicial Education Journal Volume 5 Issue 1 2022 Pages 4–5, Section IV.
43. Tomlinson, B, Torrance, AW, & Black, RW ‘ChatGPT and Works Scholarly: Best Practices and Legal Pitfalls in Writing with AI’ (2023) 76 SMU L Rev Forum 108.
44. United Nations Bangalore Principles and Commonwealth Latimer House Principles ECOSOC 2006/23.

Part VII

Volume II: SACAIR - KRR

An Override-Aware Classifier for Transparent AI

Ruvarashe Shalom Madzime^{1[[0009–0001–3469–9353]]}, Louise Leenen^{1[0000–0002–9212–550X]}, and Thomas Meyer^{2[0000–0003–2204–6969]}

¹ University of the Western Cape and CAIR, Cape Town, South Africa
4158028@myuwc.ac.za, lleenен@uwc.ac.za

² University of Cape Town and CAIR, Cape Town, South Africa
tmeyer@cair.org.za

Abstract. Many real-world decisions follow rules that hold in general but allow exceptions, such as “birds usually fly, unless they are penguins.” Most interpretable classifiers struggle to capture this pattern, leading to explanations that feel less aligned with human reasoning. This paper introduces the Defeasible Horn Classifier with Exceptions (DHCE), a symbolic model that makes this reasoning structure explicit. Each rule combines a default with its linked exceptions, so predictions can be explained step by step without relying on post-hoc tools. DHCE is learned using Answer Set Programming, which searches for globally optimal rule sets while balancing accuracy and simplicity. The resulting models consist of ranked Horn rules that provide full traceability: users can see both why a decision applies and why it may be overridden. We evaluate DHCE on standard classification benchmarks and find that it matches or outperforms leading interpretable models, a performance level that prior work shows to be competitive with classical machine learning classifiers. By making prediction decisions inherently retractable, DHCE delivers accuracy alongside explanations that mirror how people reason, making it suited for domains where understanding why a rule no longer applies is as important as the prediction itself.

Keywords: Explainable AI (XAI) · Human-aligned reasoning · Rule-based Models · Interpretable Classification · Horn rules · Transparency · Answer Set Programming

1 Introduction

Machine-learning (ML) models are increasingly being used in various high-stakes domains [1]. While these models often perform well in terms of predictive accuracy, this alone is not sufficient for real-world acceptance. It would be beneficial if these models were also interpretable, transparent, and explainable.

Explainable AI (XAI) frameworks define these qualities distinctly. A model is transparent when its internal decision-making logic is fully accessible. It is interpretable when users can follow the path from input to output and make sense of how predictions were reached. It is explainable if it can generate human-understandable justifications for its predictions, even when its internal workings

are complex or opaque [2]. Doshi-Velez and Kim [2] proposed that interpretability means being able to work through a model’s logic step by step. They highlighted simulatability (being simple enough to mentally simulate), modularity (separate parts handling specific tasks), and decomposability (where each part can be understood on its own). Rudin argued that in sensitive areas, such as healthcare or criminal justice, models should be designed to be interpretable from the start, not explained afterward with separate tools [3]. These views represent two main approaches: explanations added after training (post-hoc) and models built to be understandable from the outset (built-in interpretability).

Post-hoc methods such as LIME [4] and SHAP [5] attempt to explain the output of a trained model. They work by perturbing inputs and approximating the model’s behaviour around specific data points [3]. However, Guidotti et al. [6] note that such tools only approximate black-box behaviour. When the model is highly complex, these approximations become unreliable, hiding key parts of the actual decision process [6]. This reinforces the value of interpretable models, where the internal logic is transparent and structured in a way that can be understood without external tools.

Interpretable models are often called white-box models because their reasoning is visible. One family is decision lists: sequences of if–then rules applied top-down, where the first matching rule determines the outcome. Variants include Falling Rule Lists, which prioritize rules with higher confidence [7], and CORELS, which uses exact search to find optimal short lists [8]. A related family is decision sets, where multiple rules may apply at once, each casting a vote so that the final decision is based on an aggregate [9]. Classical rule learners such as RIPPER [10] and FOIL [11] build lists or sets greedily to reduce error. ILASP [12] provides a logic-based alternative, using Answer Set Programming (ASP) to define and search for models with user-specified structure. ASP is a declarative form of logic programming: the user states rules a solution must satisfy, and the solver automatically explores all “answer sets” that meet them [13].

Both decision lists and decision sets are explicit, but they struggle to capture how humans often think: through defaults and exceptions. A classic example is “birds normally fly, unless they are penguins.” In decision lists this requires placing the exception above the default as a separate rule, breaking their natural connection.

Another class of interpretable models learns a single Boolean expression to capture the full decision process. The Short Boolean Formula (SBF) approach [14] uses ASP to find the smallest expression (using \neg , \wedge , and \vee) that correctly classifies the data. These formulas are compact and competitive with standard classifiers such as Naïve Bayes [15]. A related approach learns tractable logical circuits, such as Ordered Binary Decision Diagrams (OBDDs) or Sentential Decision Diagrams (SDDs), which are equivalent to the classifier and support exact reasoning [16, 17].

Yet Boolean formulas and circuits often bury negations and overrides deep within the structure, without an explicit link between defaults and exceptions. They lack a clear notion of rule priority, which makes decisions harder to interpret.

In summary, post-hoc methods provide approximations of black-box models but reveal little about their true logic. Rule-based systems are transparent but struggle with exceptions. Boolean formulas and circuits achieve conciseness, yet hide reasoning in cases of conflict. None of these approaches support a natural “normally... unless...” narrative, where defaults and exceptions are directly connected. This gap limits interpretability in domains where understanding why a decision was made is as important as the outcome.

Human reasoning, by contrast, often follows a pattern of defaults and exceptions: accept what is normally true and withdraw that conclusion when an exception appears (“birds fly, unless they are penguins”). Formally, this is called defeasible reasoning. Kraus, Lehmann, and Magidor (KLM) [18] gave a formal statement of this idea: a defeasible rule says “if A then B ” in ordinary cases, but allows later information to block the inference. Because new facts can overturn prior conclusions, these logics are non-monotonic.

Classic frameworks capture this intuition. Reiter’s Default Logic [19] starts from facts, tentatively adds defaults as long as they remain consistent, and produces belief sets called extensions. Nute’s Defeasible Logic [20] separates rules into strict (always valid) and defeasible (normally valid), resolving conflicts through explicit priorities. McCarthy’s Circumscription [21] models normality by minimising “abnormal” predicates so that everything is assumed normal unless stated otherwise.

This paper introduces the Defeasible Horn Classifier with Exceptions (DHCE), an interpretable model that applies this “normally... unless...” pattern to classification tasks. DHCE is a learned rule model that captures defaults and exceptions without committing to a specific non-monotonic proof theory. Each rule contains both the default and its exception, producing a concise narrative: apply the first matching default, then check its attached exceptions. This yields step-by-step explanations that ordinary readers can follow.

Unlike rule lists that scatter exceptions or Boolean formulas that hide them in nested logic, DHCE keeps each default paired with its exception, ensuring clarity in every decision trace. On standard binary datasets, DHCE outperformed other interpretable models and reached accuracy levels comparable to classical ML methods, while still producing complete explanations for every prediction.

This paper makes three contributions: (1) a new defeasible model class with explicit exception handling, (2) an ASP encoding that finds globally optimal rule sets, and (3) empirical evidence that DHCE achieves competitive accuracy among interpretable models. Section 2 gives a motivating example, Section 3 formally defines the framework, Section 4 details the ASP encoding, Section 5 presents results, Section 6 provides a case study, and Section 7 concludes with discussion.

2 Motivating Example and Problem Statement

Consider a foundation that awards scholarships to students who either show academic merit, indicated by a high GPA (at least 3.7) and extensive community

service (at least 100 hours), or demonstrate financial need, reflected by low household income and first-generation college status. However, both paths to qualification are blocked if the student has a record of serious disciplinary misconduct or a dishonesty flag such as plagiarism. Let $Q = 1$ denote qualification, and $Q = 0$ rejection. The binary features recorded for each student are: G for high GPA, S for community service, I for low income, F for first-generation status, D for disciplinary record, and P for plagiarism. A value of 1 indicates the feature is True.

An interpretable classifier of this policy can be written as follows:

$$\begin{aligned} Q &\leftarrow G \wedge S && \text{(merit default)} \\ \neg Q &\leftarrow G \wedge S \wedge D && \text{(disciplinary exception)} \\ Q &\leftarrow I \wedge F && \text{(financial-need default)} \\ \neg Q &\leftarrow I \wedge F \wedge P && \text{(plagiarism exception)} \end{aligned}$$
(1) (2) (3) (4)

Each rule has the form $Head \leftarrow Body_1 \wedge \dots \wedge Body_k$. The symbol “ \leftarrow ” denotes implication: if all body literals are true, the head literal must also be true. The connective “ \wedge ” denotes logical conjunction, and “ \neg ” is classical negation. Thus Rule 1 says that a student with a high GPA (G) AND extensive service (S) qualifies (Q); Rule 2 overrides that by implying rejection when serious misconduct (D) is also present; Rule 3 grants qualification for low income (I) together with first-generation status (F); and Rule 4 blocks that route when plagiarism (P) is detected.

Suppose an applicant, *Emily*, has the following feature values: $G = 1$, $S = 1$, $D = 1$, $I = 0$, $F = 0$, $P = 0$. She meets the merit criteria (1), but is overruled by the disciplinary exception (2), and is therefore rejected.

One symbolic option is to learn a single Boolean formula (SBF) built only from AND (\wedge), OR (\vee) and NOT (\neg) operators that map the features to a label [14]. Here disciplinary misconduct D blocks only the merit route (Rules 1 and 2), and plagiarism P blocks only the financial-need route (Rules 3 and 4). Accordingly, an equivalent SBF encodes each default with its own blocker:

$$(G \wedge S \wedge \neg D) \vee (I \wedge F \wedge \neg P) \quad (5)$$

Substituting Emily’s values into (5) yields $(G \wedge S \wedge \neg D) = (1 \wedge 1 \wedge 0) = 0$ and $(I \wedge F \wedge \neg P) = (0 \wedge 0 \wedge 1) = 0$, so $Q = 0 \vee 0 = 0$. This classifies Emily correctly as $Q = 0$, but the reason is hidden: there is no explicit default, exception label, or clear trace of the override. Standard black box ML classifiers such as neural networks may also output $Q = 0$, but provide no explanation.

This scenario calls for a model that reflects how people naturally reason: “Normally merit qualifies, unless there’s misconduct.” DHCE captures this structure directly. For Emily, a DHCE model proceeds as follows:

1. It applies the merit default ($G \wedge S$), setting $Q = 1$.
2. It then finds the exception (D) holds, overriding the default.
3. It returns $Q = 0$ with a transparent explanation.

The result is a clear trace: “Merit default matched; disciplinary exception applied; outcome flipped to $Q = 0$.” Unlike rule lists that separate exceptions, or single formulas that could hide them, DHCE provides a clear default-and-exception narrative. Section 3 formalises this framework.

3 Formal Framework for DHCE

This section looks at the structure of DHCE, an interpretable classifier that combines default rules with attached exceptions. It works on datasets where every instance is described by binary input features (0 or 1) and a binary label. The goal is to learn a rule set that mirrors how people reason about what *normally* holds and when an *exception* overrides it.

An atom is a basic, single-feature statement, for example `GPA_high`. A positive literal is simply an atom, a negative literal is the same atom preceded by classical negation, such as `¬GPA_high`. In the datasets used to experiment on DHCE, each binary feature becomes an atom. If the feature’s value is 1 we write the positive literal `GPA_high`; if the value is 0 we write the negative literal `¬GPA_high`.

A Horn rule is a clause that takes the shape $t \leftarrow B_1, B_2, \dots, B_k$ where the head t is a single positive atom (the conclusion) and each body literal B_i is either a positive or a negative literal (a condition). Horn rules avoid nested AND/OR/NOT structures, so they read naturally as “*if all these conditions hold, then conclude t*”. Strictly restricting every rule within DHCE to be in Horn rule format has three practical benefits:

1. It is traceable: each prediction comes from one rule, optionally overridden by its local exception.
2. Placing the exception right next to its default makes the “normally … unless …” reasoning explicit.
3. Their very simple syntax keeps the search space small for the ASP encoding used by DHCE (see Section 4).

Returning to the scholarship case in Section 2, DHCE would express the logic as:

$$t \leftarrow \text{GPA_high, Service} \quad (\text{default}) \tag{6}$$

$$\neg t \leftarrow \text{GPA_high, Service, Disciplinary} \quad (\text{exception}) \tag{7}$$

This clearly shows: a student with strong merit normally qualifies, unless disqualified by misconduct. The logic is transparent, the override is explicit, and the explanation mirrors how a human reasons.

Before the DHCE system can learn rules from the dataset, the dataset has to be converted into propositional form. In this form all features describing each example must be expressed using basic true/false values. To do this, every column in the dataset is turned into a set of binary features. For numeric values like GPA or age, this is done by using the median as a cut-off point and then creating a

new feature that is true if the original value is above that threshold and false otherwise. We adopt median thresholding for numeric features to stay aligned with the SBF baseline, which Booleanises its datasets in the same way; this preserves an apples-to-apples comparison while keeping the representation simple and label-agnostic. For example, a feature `GPA_high` might be true if the GPA is above 3.7. For categorical features, a technique called one-hot encoding is used, which means that each possible category gets its own feature set to 1 if it applies, and 0 otherwise. After this step the table contains only 0s and 1s.

A DHCE model works by learning a ranked list of logical Horn rules that describe patterns within the data. The simplest kind of rule is a default rule, which describes what normally holds true (see Table 1). This means that $t = 1$ (the positive label) is predicted whenever all the conditions B_1 through B_k hold for a given instance.

However, there are often exceptions to such general patterns. An exception rule is used to override a default when more specific information is available (Table 1). This says that even if the default conditions B_1 through B_k hold, the conclusion $t = 1$ is cancelled if extra conditions E_1 through E_m also hold. In that case, the prediction flips to $t = 0$. Multiple exceptions per default are allowed in our encoding. The number is bounded by the constant (`maxE`), which we set during the model search. Each exception is evaluated only when its parent default matches, since exceptions are attached to that default. We do not allow exceptions to exceptions in the current design. This choice keeps every override local to its default and preserves a short, readable trace where each default appears next to its own exception.

After evaluating all defaults and their attached exceptions, the DHCE model introduces two final rule layers to guarantee full coverage of the dataset. First comes a fallback rule, a single-literal Horn rule added when no previous default or exception applies (Table 1). Here, F is the single literal that best classifies the remaining uncovered training examples. The selection is made by scoring each available literal against the uncovered cases and choosing the one that minimises misclassification. Like defaults, the fallback rule may include exceptions that cancel its effect. If even the fallback does not match, the catch-all rule fires. The catch-all has an empty body (Table 1); it always fires and predicts the majority class in the training data. For example, if most training examples have label $t = 1$, then the catch-all predicts $t = 1$ for any unmatched instance. The two layers are kept distinct by design to preserve traceability and clear tie-breaking.

Therefore, each instance can be traced back to a specific rule layer: default, exception, fallback, or catch-all, providing users with a transparent trace that is not available in black-box models.

DHCE classifies each instance by applying its rules in a clear, step-by-step order, using built-in priority handling to decide which rule should apply first. The model is defeasible by nature: conclusions can be withdrawn if stronger evidence appears. Rules are assigned numeric priorities to control conflict resolution. Learned default rules occupy the top levels, with ascending values (e.g., 0, 10, 20, ...), where lower numbers indicate higher priority. Each exception inherits its

Table 1. Formulas used in DHCE rules

Rule Type	Formula
Default	$t \leftarrow B_1, B_2, \dots, B_k$
Exception	$\neg t \leftarrow B_1, B_2, \dots, B_k, E_1, \dots, E_m$
Fallback	$t \leftarrow F$
Catch-all	$t \leftarrow$

parent default’s priority, ensuring they are evaluated together. The fallback rule and its possible exception are placed just below all defaults, and the catch-all rule sits at the very bottom. When an input is received, DHCE evaluates all applicable rules and selects the one with the lowest priority number. If multiple rules share this priority, the system chooses the rule with the most conditions in its body, favoring specificity.

This setup means that exceptions only apply when their parent default matches, and they always win in a conflict because their body strictly contains the default’s body. Keeping exceptions at the same priority as their parent default avoids extra complexity and keeps the reasoning path local and easy to follow. The evaluation unfolds in a fixed order: scan defaults, apply the first match, check for an exception, otherwise move to the fallback, and finally the catch-all. This process guarantees that each instance receives exactly one label. Because each prediction is backed by a rule with a known priority and body, users can always trace why one rule overruled another.

To learn DHCE models, ASP is used. ASP is a declarative framework for hard search problems. A program states rules and constraints that any solution must satisfy, and a solver searches for stable models that meet those conditions. ASP is widely used in knowledge representation and reasoning [13]. Clingo is the standard toolchain. It grounds the program to a propositional form and then calls clasp. Clasp is the solver that performs the search and optimisation [22].

In this work ASP generates candidate DHCE rule sets, checks structural constraints such as uniqueness and ordering, and optimises for accuracy and simplicity. The resulting models classify the training data while keeping explanations readable.

Concretely, clingo grounds the DHCE template and hands it to clasp, which searches for stable models under a lexicographic objective. The encoding proposes defaults up to `maxD`, and for each default up to `maxE` attached exceptions, with each body limited by `maxBody`. For each candidate rule set the program simulates classification on the training rows to count errors and records simple size measures. Optimisation proceeds in levels: minimise training error first, then prefer fewer defaults, fewer exceptions, and shorter bodies. During model selection a grid over `maxD`, `maxE`, and `maxBody` is scanned. The configuration with the lowest test error is chosen, breaking ties by total literal count.

Consider the following example. Suppose $\text{maxD} = 1$, $\text{maxE} = 2$, $\text{maxBody} = 2$. A candidate default $t \leftarrow a_1, a_2$ is proposed. Two exceptions can attach to this default, for example $\neg t \leftarrow a_1, a_2, e_1$ and $\neg t \leftarrow a_1, a_2, e_2$. The solver tests these

candidates on the training rows. If the default reduces error and some mistakes occur only when e_1 holds, the first exception is kept; if a distinct set of mistakes occurs with e_2 , the second exception is also kept, provided both stay within the size bounds. If multiple candidates tie on error, the version with shorter bodies is preferred. A single literal fallback and an empty body catch-all are then added so that every instance receives a label. This illustrates that more than one exception may attach to the same default while the priority and tie breaking keep the override local and readable.

To evaluate how well DHCE performs, this study adopts the same experimental setup used by Jaakkola et al. [14], who introduced the SBF method. Like DHCE, their method uses ASP to learn interpretable models from binary (true/false) data. Both systems are evaluated using ten random 50/50 splits of each dataset, half the data for training, and half for testing. The goal is to compare the training and test error across different splits and see how DHCE performs relative to SBF.

The prediction error for a learned rule set \mathcal{R} is the proportion of examples on which its prediction $\hat{y}\mathcal{R}(x)$ disagrees with the true label $y(x)$. We write S for the training set and T for the test set (with sizes $|S|$ and $|T|$), let x range over the instances in either set, and use the indicator $\mathbf{1}[\cdot]$ that returns 1 when its argument is true and 0 otherwise. With this notation the training error and test error rates are

$$\text{Errortrain}(\mathcal{R}) = \frac{1}{|S|} \sum_{x \in S} \mathbf{1}[\hat{y}\mathcal{R}(x) \neq y(x)], \quad (8)$$

$$\text{Errortest}(\mathcal{R}) = \frac{1}{|T|} \sum_{x \in T} \mathbf{1}[\hat{y}\mathcal{R}(x) \neq y(x)]. \quad (9)$$

These error values are used to find the best rule set for each split. All results are stored inside the ASP generated answer sets for traceability.

This lexicographic optimisation ensures that the final model is not only accurate but also as simple and interpretable as possible. Because DHCE and SBF both use ASP, the same kind of data, and the same evaluation process, they are compatible for a comparison. It allows for directly assessing whether adding defeasible rule structure improves interpretability without affecting performance.

4 Implementation and Experimental Setup

SBF has been shown to compete with Naïve Bayes and decision trees [14]. DHCE seeks similar or better predictive accuracy while delivering more interpretable classifications. During evaluation, the system logs training error, test error, total rule set size, and per layer coverage, which are analysed in Section 5.

This section explains how DHCE was implemented, how it was evaluated, and why certain design decisions were made. All three datasets used in this study are publicly available from the UCI Machine Learning Repository [23], and they were the same datasets used in the SBF study [14]. The Breast Cancer Wisconsin (Original) dataset predicts whether a tumor is benign or malignant

based on microscopic cytology features. The Statlog (German Credit) dataset assesses whether a loan applicant is a good or bad credit risk, using financial and demographic attributes. The Ionosphere dataset classifies radar signal returns as “good” or “bad” for identifying ionospheric structures.

Each dataset was preprocessed to convert it into Boolean (true/false) form. As mentioned before, for numeric features, median thresholding was used and for categorical features, one-hot encoding was used. After this transformation, every data point is represented as a flat vector of 0s and 1s; concretely: Breast-Cancer has 683 rows with 10 Boolean attributes (cytology: benign/malignant), German-Credit has 1000 rows with 21 Boolean attributes (credit risk: good/bad), and Ionosphere has 351 rows with 35 Boolean attributes (radar signal: good/bad).

In the DHCE implementation, each example is encoded in ASP as a set of `val/3` atoms: `val(RowID, a(Attribute, Category?), 0/1)`. This means that the feature ‘`a(Attribute, Category)`’ has value 0 or 1 for a given row (data point). The optional `Category` is only used when the feature comes from a one-hot encoded categorical variable, as in the German-Credit dataset. For example, `val(7, a(3,2), 1)`. means that row 7 has category 2 active for attribute 3. In contrast, datasets like Breast-Cancer and Ionosphere use simpler encodings like `a(5)` for attribute 5.

Each dataset ends with a special column for the target label, this becomes the attribute to be predicted in DHCE. The exact target label positions are: `a(10)` for Breast Cancer, `a(21)` for German-Credit, and `a(35)` for Ionosphere.

To ensure a fair comparison with SBF, the evaluation protocol described by Jaakkola *et al.* [14] was followed. Each dataset was split into ten random 50/50 training–test partitions using fixed seeds. A driver script handles this protocol. It writes each split into grounded ASP facts, runs the solver with various parameter settings, and logs the results. For each split, the best model is selected based on test error. If several configurations tie on error, the system chooses the one with the fewest literals. This evaluation method aligns with SBF’s evaluation method to support an apples-to-apples comparison.

The DHCE ASP encoding has three tasks: it constructs a layered Horn rule list (defaults with attached exceptions, a single literal fallback, and an empty body), it simulates those rules on the training rows to count errors, and it exposes error and size summaries that the optimiser can minimise. Concretely, the encoding proposes bodies for learned defaults up to `maxD`; for each default it allows up to `maxE` attached exceptions, and all bodies obey `maxBody`. The fallback slot is constrained to exactly one literal and the slot to an empty body. This produces a compact, ranked structure where overrides stay local to the default they qualify.

Misclassification is recorded whenever the predicted label differs from the ground truth for a training row. Optimisation is lexicographic: minimise training errors first; among ties prefer fewer defaults; then fewer exceptions; then the sum of default body lengths; then the sum of exception body lengths. The label index is dataset specific as described earlier in this section (`a(10)` for Breast Cancer, `a(21)` for German Credit, `a(35)` for Ionosphere). This summary keeps the learning

description self contained while making clear how DHCE differs from flat formula classifiers such as SBF.

To control the complexity of the models it learns, DHCE searches over a small grid of rule-size limits. This grid defines three aspects of model structure: the maximum number of default rules (`maxD`), the number of exceptions allowed per default (`maxE`), and the number of literals allowed in the body of any rule (`maxBody`). The solver explores all combinations of these values to generate a range of candidate rule sets. The grid used in this study includes $\text{maxD} \in \{3, 4\}$, $\text{maxE} \in \{1, 2\}$, and $\text{maxBody} \in \{1, 2, 3, 4\}$, which was found to offer a good trade-off between flexibility and runtime in pilot experiments.

For each train–test split, the solver evaluates every rule-size setting from the grid and selects the model with the lowest test error. If multiple models achieve the same accuracy, the one with the fewest total literals is preferred. This helps prevent overfitting by favouring simpler rule sets when possible.

Internally, the ASP encoding further enforces regularisation by disallowing redundant exceptions and ensuring that each rule covers at least one training instance not already handled by higher-priority rules. These structural constraints guide the solver toward non-redundant models.

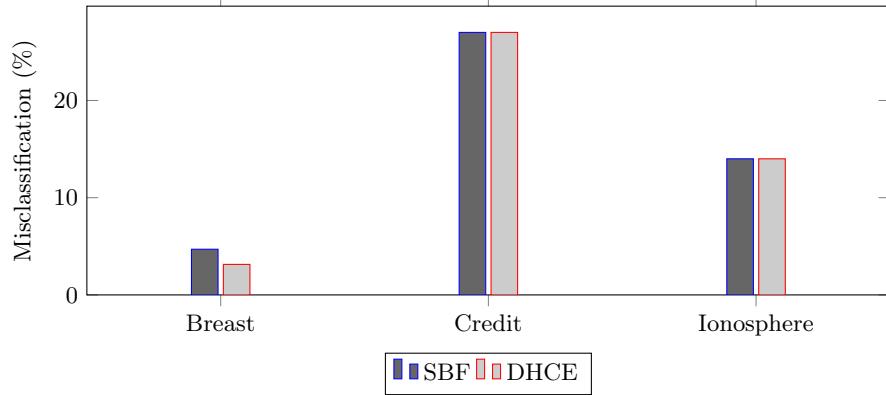
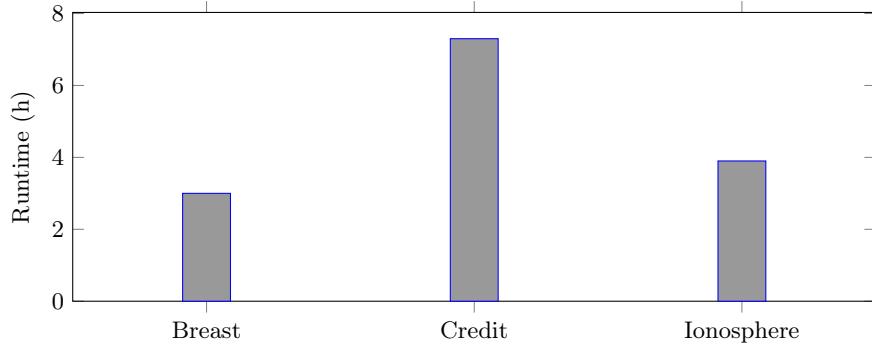
SBF has already been shown to compete with traditional classifiers such as Naïve Bayes and decision trees [14]. DHCE aimed to achieve similar or better predictive accuracy while delivering more interpretable, classifications. During evaluation, the system logged key metrics, including training error, test error, total rule set size, and per-layer rule coverage. This was all saved and used for analysis in Section 5.

5 Results

The evaluation in this section builds directly on the protocol introduced in Section 3 and implemented as described in Section 4. As outlined earlier, each benchmark dataset was partitioned into ten fixed 50/50 train–test splits. For each split, DHCE searched over a grid of rule structures and selected the rule set that achieved the lowest test error, breaking ties using the smallest total number of literals. This protocol matches the setup used by Jaakkola et al. [14], enabling a direct comparison with SBF. The results presented next, summarise the characteristics of the best models discovered.

Figure 1 reports the lowest test error across all splits for each dataset. DHCE outperforms SBF on Breast-Cancer, reducing error by 1.5 percentage points. On German-Credit and Ionosphere, both models achieve identical error rates. These results confirm that the addition of defeasible rule layers does not degrade predictive accuracy, and in some cases can improve it.

When computing the DHCE rule sets, no strict timeout was imposed. Instead, the solver was allowed to run until it found an optimal rule set for each best split. This ensured that all models used in evaluation were globally optimal with respect to the specified objective. Figure 2 shows the total runtime per dataset. The German-Credit dataset required the longest solving time (7.3 hours), which

**Fig. 1.** Best-split misclassification rate.**Fig. 2.** Runtime optimal DHCE model (best split).

aligns with its size of 1 000 rows and 21 Boolean attributes. In contrast, the Breast-Cancer dataset (683 rows, 9 attributes) was solved in 3.0 hours, while Ionosphere (351 rows, 34 attributes) took 3.9 hours. These differences suggest that both dataset size and attribute dimensionality affect solving time. Across all datasets, the average runtime was approximately 4.7 hours. These times are reasonable for offline rule discovery, especially given the interpretability and structure DHCE provides in return.

The winning SBF rules are compact Boolean expressions, summarised per dataset in Table 2 [14]. These rules are notably short, 3 to 4 literals each. However, their simplicity comes at the cost of interpretability: the entire decision logic is compressed into a single, flat formula, encoding all reasoning inside nested logical operations. In contrast, Table 3 presents the explicit DHCE rule sets that match or exceed the accuracy of SBF. While DHCE rules use more literals overall: 7 for German-Credit and 9 for both Breast-Cancer and Ionosphere, the increase in size enables significantly clearer logic decomposition. Each rule is assigned

to a named layer (Default, Exception, Fallback, Catch-all), offering a faithful “normally . . . unless . . .” structure that mirrors human decision-making. These named layers allow users to see exactly when a rule applies and what specific exception (if any) overrides it.

Table 2. Best-performing SBF formulas per dataset.

Dataset	SBF Formula
Breast-Cancer (4 lit.)	$\neg((a[1] \wedge a[6]) \vee a[5]) \wedge a[3]$
German-Credit (3 lit.)	$\neg(a[1, 1] \wedge a[2]) \vee a[17, 4]$
Ionosphere (4 lit.)	$((a[8] \wedge a[12]) \vee a[15]) \wedge a[1]$

The SBF formulas in Table 2 are short, but they combine all logic into a single compact expression, which can be difficult to read. For example, the Breast-Cancer rule uses nested operations like $\neg((a[1] \wedge a[6]) \vee a[5]) \wedge a[3]$, which the user must unpack for understanding. In contrast, the DHCE rule sets in Table 3 break down the logic into separate layers: defaults, exceptions, fallback, and catch-all. This separation makes it easier to see what normally holds, when it is overridden, and how each case is handled. Although DHCE uses more literals overall, this extra detail improves clarity and helps explain each prediction step by step.

Table 3. DHCE best-split rule sets across datasets

Dataset	Rule Type	Rule	Exception(s)
CANCER (9 literals)	Default 1	$1 \leftarrow a(2, 0)$	$0 \leftarrow a(2, 0) \wedge a(8, 1)$
	Default 2	$1 \leftarrow a(7, 0)$	$0 \leftarrow a(7, 0) \wedge a(8, 1)$
	Fallback	$1 \leftarrow a(1, 1)$	$0 \leftarrow a(1, 1) \wedge a(3, 1)$
	Catch-all	majority label 1	—
CREDIT (7 literals)	Default 1	$1 \leftarrow a(16, 4) = 1$	$0 \leftarrow a(16, 4) = 1 \wedge a(4, 10) = 1$
	Fallback	$1 \leftarrow a(16, 4) = 0$	$0 \leftarrow a(16, 4) = 1 \wedge a(4, 8) = 1$
	Catch-all	majority class 1	—
	Default 1	$1 \leftarrow a(1, 0)$	$0 \leftarrow a(1, 0) \wedge a(15, 1)$
IONOSPHERE (9 literals)	Default 2	$1 \leftarrow a(12, 0)$	$0 \leftarrow a(12, 0) \wedge a(33, 0)$
	Fallback	$1 \leftarrow a(1, 1)$	$0 \leftarrow a(1, 1) \wedge a(8, 0)$
	Catch-all	majority label 1	—

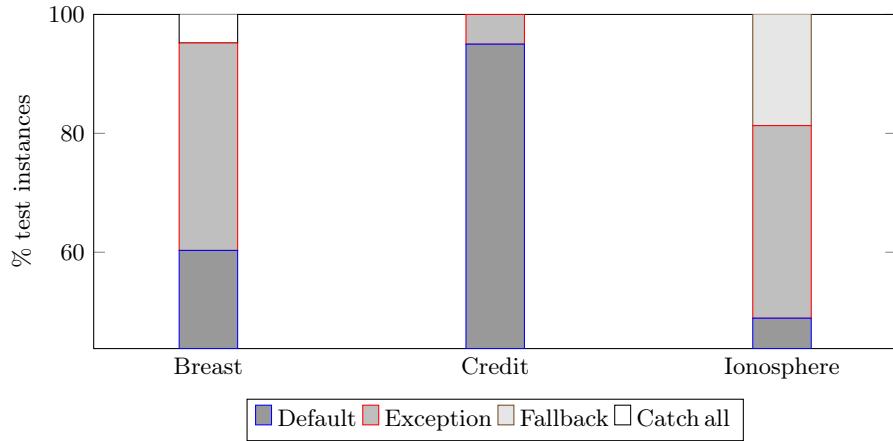


Fig. 3. Layer coverage of the *best* DHCE rule set on each dataset.

Figure 3 shows for the Breast-Cancer dataset how DHCE explains predictions using Default rules 60.3% of the time and Exception rules 34.9%, with only 4.9% handled by the Catch-all. In Ionosphere, 48.9% of test instances are covered by Defaults, 32.4% by Exceptions, and 18.8% by the Fallback rule—again, no reliance on Catch-all. On German-Credit, a dominant Default rule covers 95.0% of predictions, with only 5.0% requiring an Exception. DHCE uniquely provides this layered breakdown, providing transparency for the user.

DHCE shows strong overall performance across all benchmarks. In terms of accuracy, it matches or exceeds SBF performance, notably achieving a 1.54 percentage-point improvement on the Breast-Cancer dataset in Figure 1. More importantly, DHCE offers an advantage in interpretability: its rules are structured explicitly using “normally … unless …” clauses (see Table 3), making the reasoning process transparent and easy to follow, unlike Boolean expressions used by SBF. From a classifier-complexity perspective, DHCE maintains typically 7–9 literals while remaining practical in terms of runtime. As shown in Figure 2, model training takes between 3 and 7.3 hours, making the method feasible for applications where interpretability is a priority.

6 A Case Study

This section shows how DHCE mirrors human reasoning more closely than a single Boolean formula, even when both models reach the correct prediction. For this example, we consider a Breast-Cancer patient whose features activate both a default and an attached exception in the DHCE rule set.

To give context, Table 4 lists the ten columns that appear in the Breast-Cancer dataset. The first nine are input attributes extracted from microscopic images and then binarised at the median: a value of 0 means the measurement falls in the lower half of the training data, and 1 means it lies in the upper half [23]. The

last column, $a[10]$, is the *target label* supplied by the pathologist, in every model used here a label of 0 is interpreted as *benign* and 1 as *malignant*.

Using the feature definitions given in Table 4, consider a patient with binarised cytology profile (CT, UCS, UCSh, MA, SECS, BN, BC, NN, MIT) = (0, 0, 0, 0, 0, 0, 0, 1, 0). This individual has $a[2] = 0$ (cells are uniformly sized). However, prominent nucleoli are also present ($a[8] = 1$), a known override feature. These conditions trigger both a default and its linked exception in the DHCE rule set (see Table 3). Default 1 fires (see Table 3): $1 \leftarrow a[2]=0$ “Having uniform cell size normally leads to a prediction of malignant.” Exception overrides: $0 \leftarrow a[2]=0 \wedge a[8]=1$ “Having uniform cell size normally leads to a malignant prediction, unless prominent nucleoli are also present, in which case the prediction is benign.” Final DHCE prediction: benign (0). SBF also predicts class 0, using its learned 4-literal formula: $\neg((a[1] \wedge a[6]) \vee a[5]) \wedge a[3]$. Substituting these values yields: $\neg((0 \wedge 0) \vee 0) \wedge 0 = 0$, which matches DHCE’s output. However, the SBF formula does not show clearly how this decision was made. It is difficult to tell which features led to the result, and there is no indication that one rule may have been overridden by another. All of the reasoning is packed into a single expression, which hides the steps that were followed. In contrast, DHCE shows its reasoning step by step. It clearly indicates that a default rule first predicted a malignant tumour, but this was then changed because an exception applied. This kind of explanation is easier to follow and matches how people usually reason.

This example also reflects what happens more broadly. On the full Breast-Cancer test set, DHCE made fewer errors than SBF, improving accuracy by 1.5 percentage points (Figure 1). It also gave a clear rule-layer explanation for all test cases, showing whether the result came from a default, an exception, or the catch-all rule (see Table 3). Even when both models give the same answer, DHCE gives a better decision trace by showing both what normally happens and when that rule no longer applies.

Table 4. Binarised Breast-Cancer columns ($a[1] - a[9]$ indicate 0 = below median, 1 = above; $a[10]$ is the diagnosis.)

Index Description		0 = Lower half 1 = Upper half	
$a[1]$	Clump Thickness (CT)	Thin clumps	Dense clumps
$a[2]$	Uniformity Cell Size (UCS)	Uniform cells	Size variation
$a[3]$	Uniformity Cell Shape (UCSh)	Regular shapes	Irregular shapes
$a[4]$	Marginal Adhesion (MA)	Weak adhesion	Strong bonding
$a[5]$	Single Epithelial Cell Size	Small cells	Large cells
$a[6]$	Bare Nuclei (BN)	Few nuclei	Many nuclei
$a[7]$	Bland Chromatin (BC)	Fine texture	Coarse texture
$a[8]$	Normal Nucleoli (NN)	Not visible	Prominent
$a[9]$	Mitoses (MIT)	Low activity	Elevated activity
$a[10]$	Diagnosis (label)	Benign	Malignant

7 Discussion and Conclusion

The experimental results confirm that DHCE achieves competitive accuracy while preserving its transparent “*normally ... unless ...*” structure. On the Breast-Cancer dataset [23] it reduced error from 4.7% to 3.14%, and on German-Credit and Ionosphere [23] it matched the 27.0% and 14.0% baselines, showing that attaching an explicit exception to every default does not compromise performance (See Figure 1). DHCE demonstrates how a model built from defaults and exceptions can tell the user both *why* a default fired and *when* it was overridden.

Overfitting is controlled by keeping small bounds on maxD, maxE, and maxBody. When accuracy ties, selection prefers fewer rules and shorter bodies, which acts as regularisation. The encoding prevents redundancy by requiring each new default to cover at least one previously uncovered instance and by disallowing duplicate exception bodies. Finally, models are chosen by the lowest test error over repeated splits, so large train-test gaps are penalised.

Several related studies also fuse explanations with defeasible reasoning. Brewka and Ulbricht’s work on *strong explanations* shows how to identify the minimal facts that guarantee a conclusion despite exceptions [24]. Everett *et al.* [25] extend propositional KLM logic [18] with an explanation layer that highlights the rules ensuring a conclusion in the most normal situation. Chama *et al.* [26] adapt this idea to description logic, ranking worlds by normality and returning the smallest set of rules and exceptions that preserve a conclusion in the top-ranked worlds. Rienstra [27] applies similar ideas to Boolean classifiers with background knowledge, refining Darwiche and Hirth’s monotonic “sufficient reasons” [28] into “ideal reasons” that are concise, knowledge-aware, and free of redundancy.

These approaches differ in technique but share a single goal: tell users not only *what* the system decided but also *why* it stands and *when* it would change. DHCE pursues this same spirit from a learning perspective. Instead of adding an explanation layer, DHCE learns rules that already combine defaults with their local exceptions, so every prediction comes with a trace. In DHCE, the structure itself—defaults, exceptions, and their priorities—forms the explanation.

At present, DHCE handles only binary features and labels. A natural next step is to extend the default–exception scheme to multi-class problems and numeric or ordinal attributes. Training currently relies on Clingo, which guarantees optimal rule sets but scales poorly as feature counts grow. Future research into incremental grounding, heuristic pruning, or anytime search could improve runtime for larger datasets. Another direction is to compare DHCE directly with post-hoc tools like LIME or SHAP [4], clarifying its practical advantages. Finally, embedding DHCE in an explicit non-monotonic framework could provide formal guarantees that further support its use in real applications.

Acknowledgements

This work is based on the research supported in part by the National Research Foundation of South Africa (REFERENCE NO: SAI240823262612).

References

- [1] Andreas Fuster, Paul Goldsmith-Pinkham, Tarun Ramadorai, and Ansgar Walther. "Predictably Unequal? The Effects of Machine Learning on Credit Markets". In: *The Journal of Finance* 77.1 (2022), pp. 5–47. DOI: 10.1111/jofi.13090.
- [2] Finale Doshi-Velez and Been Kim. *Towards a Rigorous Science of Interpretable Machine Learning*. arXiv preprint arXiv:1702.08608. 2017. DOI: 10.48550/arXiv.1702.08608.
- [3] Cynthia Rudin. *Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead*. arXiv preprint arXiv:1811.10154. 2019. DOI: 10.48550/arXiv.1811.10154.
- [4] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "'Why Should I Trust You?': Explaining the Predictions of Any Classifier". In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2016, pp. 1135–1144. DOI: 10.1145/2939672.2939778.
- [5] Scott Lundberg and Su-In Lee. *A Unified Approach to Interpreting Model Predictions*. arXiv preprint arXiv:1705.07874. 2017. DOI: 10.48550/arXiv.1705.07874.
- [6] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. "A Survey of Methods for Explaining Black Box Models". In: *ACM Computing Surveys* 51.5 (2019). DOI: 10.1145/3236009.
- [7] Fulton Wang and Cynthia Rudin. "Falling Rule Lists". In: *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*. Vol. 38. Proceedings of Machine Learning Research. PMLR, 2015, pp. 1013–1022.
- [8] Elaine Angelino, Nicholas Larus-Stone, Daniel Alabi, Margo Seltzer, and Cynthia Rudin. "Learning Certifiably Optimal Rule Lists for Categorical Data". In: *Journal of Machine Learning Research* 18 (2018), pp. 1–78.
- [9] Himabindu Lakkaraju, Stephen H. Bach, and Jure Leskovec. "Interpretable Decision Sets: A Joint Framework for Description and Prediction". In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2016, pp. 1675–1684. DOI: 10.1145/2939672.2939874.
- [10] William W. Cohen. "Fast Effective Rule Induction". In: *Proceedings of the Twelfth International Conference on Machine Learning*. Morgan Kaufmann Publishers Inc., 1995, pp. 115–123.
- [11] J. Ross Quinlan and R. Michael Cameron-Jones. "FOIL: A Midterm Report". In: *European Conference on Machine Learning*. Springer, 1993, pp. 3–20.
- [12] Mark Law, Alessandra Russo, and Krysia Broda. *The ILASP System for Inductive Learning of Answer Set Programs*. arXiv preprint arXiv:2005.00904. 2020. DOI: 10.48550/arXiv.2005.00904.
- [13] Gerhard Brewka, Thomas Eiter, and Mirosław Truszczyński. "Answer Set Programming at a Glance". In: *Communications of the ACM* 54.12 (2011), pp. 92–103. DOI: 10.1145/2043174.2043195.

- [14] Reijo Jaakkola, Tomi Janhunen, Antti Kuusisto, Masood Feyzbakhsh Rankooh, and Miikka Vilander. “Short Boolean Formulas as Explanations in Practice”. In: *Logics in Artificial Intelligence (JELIA 2023)*. Vol. 14281. Lecture Notes in Artificial Intelligence. 2023, pp. 90–105.
- [15] Pedro Domingos and Michael Pazzani. “On the Optimality of the Simple Bayesian Classifier under Zero-One Loss”. In: *Machine Learning* 29.2–3 (1997), pp. 103–130. doi: 10.1023/A:1007413511361.
- [16] Jinbo Huang and Adnan Darwiche. “The Language of Search”. In: *Journal of Artificial Intelligence Research* 29 (2007), pp. 191–219. doi: 10.1613/jair.2097.
- [17] Adnan Darwiche. “SDD: A New Canonical Representation of Propositional Knowledge Bases”. In: *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence*. 2011, pp. 819–826. doi: 10.5591/978-1-57735-516-8/IJCAI11-143.
- [18] Sarit Kraus, Daniel Lehmann, and Menachem Magidor. “Nonmonotonic Reasoning, Preferential Models and Cumulative Logics”. In: *Artificial Intelligence* 44.1–2 (1990), pp. 167–207.
- [19] Raymond Reiter. “A Logic for Default Reasoning”. In: *Artificial Intelligence* 13.1–2 (1980), pp. 81–132. doi: 10.1016/0004-3702(80)90014-4.
- [20] Donald Nute. “Defeasible Reasoning”. In: *Handbook of Philosophical Logic*. Vol. 3. 1987, pp. 353–395.
- [21] John McCarthy. “Circumscription—A Form of Non-Monotonic Reasoning”. In: *Artificial Intelligence* 13.1–2 (1980), pp. 27–39.
- [22] Martin Gebser, Roland Kaminski, Benjamin Kaufmann, and Torsten Schaub. *Answer Set Solving in Practice*. Morgan & Claypool Publishers, 2012.
- [23] Dheeru Dua and Casey Graff. *UCI Machine Learning Repository*. 2019.
- [24] Gerhard Brewka and Markus Ulbricht. “Strong Explanations for Nonmonotonic Reasoning”. In: *Description Logic, Theory Combination, and All That*. Springer, 2019, pp. 135–146. doi: 10.1007/978-3-030-22102-7_6.
- [25] Lloyd Everett, Emily Morris, and Thomas Meyer. “Explanation for KLM-Style Defeasible Reasoning”. In: *Proceedings of the Southern African Conference for Artificial Intelligence Research*. 2021, pp. 192–207.
- [26] Victoria Chama, Steve Wang, Thomas Meyer, and Giovanni Casini. “Defeasible Justification for KLM-Style Logic”. In: *Proceedings of the 37th International Workshop on Description Logics*. 2024.
- [27] Tjitzé Rienstra. “Explaining Boolean Classifiers with Non-Monotonic Background Theories”. In: *Artificial Intelligence and Machine Learning — 35th Benelux Conference, BNAIC/Benelearn 2023, Revised Selected Papers*. Vol. 2187. Communications in Computer and Information Science. Springer, 2025, pp. 174–188. doi: 10.1007/978-3-031-74650-5_10.
- [28] Adnan Darwiche and Pierre Marquis. “On the (Complete) Reasons Behind Decisions”. In: *Proceedings of the 24th European Conference on Artificial Intelligence*. 2020, pp. 2407–2414. doi: 10.3233/FAIA200375.