CMP6200
Individual Undergraduate Project
2024 – 2025

University Artificially Intelligent Assistant

Course: Computer & Data Science
Student Name: Lewis Higgins
Student Number: 22133848
Supervisor Name: Dr. Atif Azad

Abstract

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Curabitur et erat consectetur, scelerisque eros nec, ultricies est. Donec mi ipsum, imperdiet vitae arcu quis, luctus venenatis quam. Integer ac massa a augue venenatis fringilla. Etiam posuere libero sed nulla tristique volutpat.

Acknowledgements

I would like to acknowledge. . .

Contents

Glossary

Term	Definition
RAG	Retrieval-Augmented Generation is...

List of Figures

List of Tables

Introduction

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Curabitur et erat consectetur, scelerisque eros nec, ultricies est. Donec mi ipsum, imperdiet vitae arcu quis, luctus venenatis quam. Integer ac massa a augue venenatis fringilla. Etiam posuere libero sed nulla tristique volutpat.

1.1 Problem definition

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Curabitur et erat consectetur, scelerisque eros nec, ultricies est. Donec mi ipsum, imperdiet vitae arcu quis, luctus venenatis quam. Integer ac massa a augue venenatis fringilla. Etiam posuere libero sed nulla tristique volutpat.

1.2 Scope

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Curabitur et erat consectetur, scelerisque eros nec, ultricies est. Donec mi ipsum, imperdiet vitae arcu quis, luctus venenatis quam. Integer ac massa a augue venenatis fringilla. Etiam posuere libero sed nulla tristique volutpat.

1.3 Rationale

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Curabitur et erat consectetur, scelerisque eros nec, ultricies est. Donec mi ipsum, imperdiet vitae arcu quis, luctus venenatis quam. Integer ac massa a augue venenatis fringilla. Etiam posuere libero sed nulla tristique volutpat.

1.4 Aims and Objectives

This project aims to aid new and existing students alike while they are attending university with helpful information about university itself, such as university societies, locations/campuses, and policies through the medium of a digital chatbot companion to converse with. The project's objectives are:

- Conduct a thorough literature review on the surrounding topics, namely AI, LLMs and NLP.
- Create effective documentation for all stages of development, highlighting challenges faced during the process.
- Leverage Retrieval-Augmented Generation alongside a cloud-based LLM to query a vector database of university-related data.
- Develop a chatbot capable of accurately answering user queries related to university buildings, policies, and societies with a minimum 80% accuracy rate.
- Evaluate the effectiveness of an AI assistant on university student acclimatization.

1.5 Background information

Possibly unnecessary.

Literature Review

2.1 Review of Literature

2.1.1 Artificial Intelligence (AI)

Researchers have always wanted to harness the processing power of computers to act in a manner indistinguishable from that of humans from as long ago as 1950, where the question was posed 'Can machines think?' (Turing, 1950). Ever since, constant innovations were made in computer intelligence and machine learning, from playing games of checkers at a better level than human players (Samuel, 1959) to classifying the contents of millions of images using convolutional neural networks (Krizhevsky, Sutskever and Hinton, 2012).

Recently, AI is used across many disciplines for different purposes to complete tasks faster than, and in some cases better than, human workers, especially with the introduction of large language models (LLMs) (Maedche et al., 2019). Wirtz et al. (2018) write that 'service robots' [1] can complete a variety of tangible or intangible actions, such as two-way conversation with chatbots.

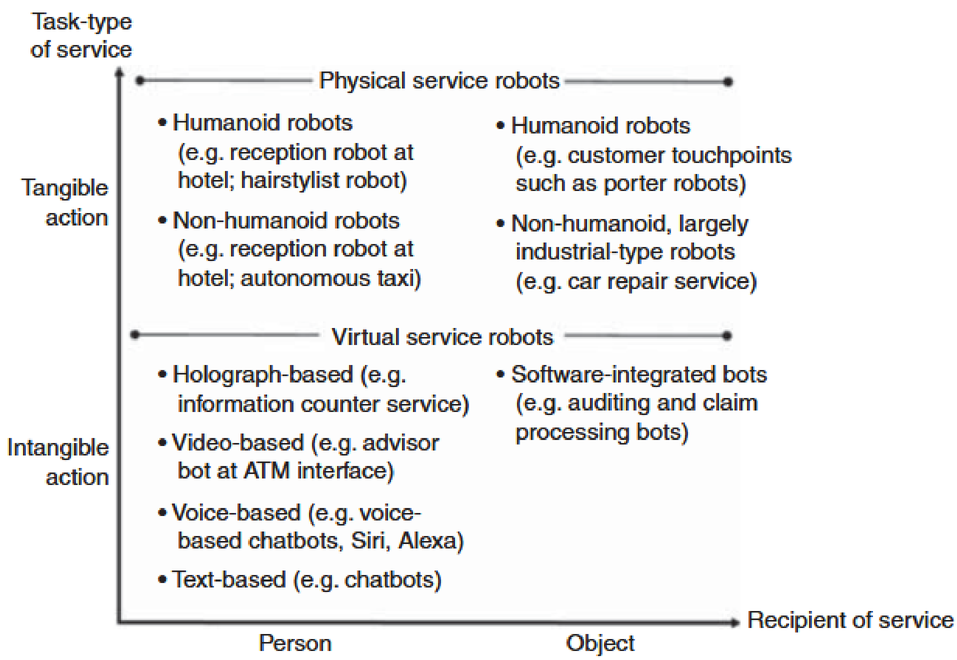

Figure 2.1: Service robots categorization by task-type and recipient of service (Wirtz et al., 2018).

When developing an AI project, it is important that the development process is ethical and human-centred, which is known as Human-Centred AI (HCAI). Another issue is the "black-box problem" - the inability to know an AI's reasoning, meaning that eXplainable AI (XAI) is a growing necessity (Miró-Nicolau, Jaume-i-Capó and Moyà-Alcover, 2025).

[1] Defined as "system-based autonomous and adaptable interfaces that interact, communicate and deliver service to an organization's customers" (Wirtz et al., 2018, p.909)

Focusing on HCAI and XAI means the focus shifts from the machine to the user and their experience using the AI. Shneiderman (2020) strongly advocates for the promotion of HCAI for the benefit of both companies and their users, which is a commonly accepted idea due to the ethical risks of using AI.

Because AI calculates outcomes from its training data rather than understanding social norms and perspectives, using it in sociotechnical systems poses serious risks due to the 'traps' it can fall into, because it cannot account for every possibility such as the personal tendencies and biases of its users (Selbst et al., 2019), and therefore developers require a shift in focus - from the final product to the development process itself and end users, which also echoes Shneiderman's views.

2.1.2 Natural language processing (NLP)

The ability for a computer to interpret and understand human language greatly enhances the scale of their capabilities. This was recognised during the 1950s, where machine translation from Russian to English was demonstrated for the first time, albeit in a basic form (Jones, 1994). Ever since, NLP has been a key topic in computing, especially in recent years, with its applications widening in scope with modern processing power.

One of the key advancements in NLP is vectorisation, a process where data is embedded into a numerical equivalent that a computer can interpret, enabling Natural Language Understanding (NLU) and the identification of semantic similarities between words through the use of an embedding model like Word2Vec (Mikolov et al., 2013) without the need to manually label data. Word2Vec was a key innovation in NLP, and Mikolov and Le went on to improve it further with Doc2Vec (Le and Mikolov, 2014), which could embed entire documents into semantically searchable vectorised forms.

Embedding models have further improved since, most notably with Vaswani et al. (2017)'s Transformer architecture enhancing models such as BERT (Devlin et al., 2019), which establishes context through analysing multiple neighbours of a word rather than reading from left to right, gaining a higher understanding of the text it processes. Many embedding models have since been developed, though one of the most reputable is OpenAI's recent text-embeddings-3 model (OpenAI, 2024c), which can be used in the development of the chatbot at a low cost.

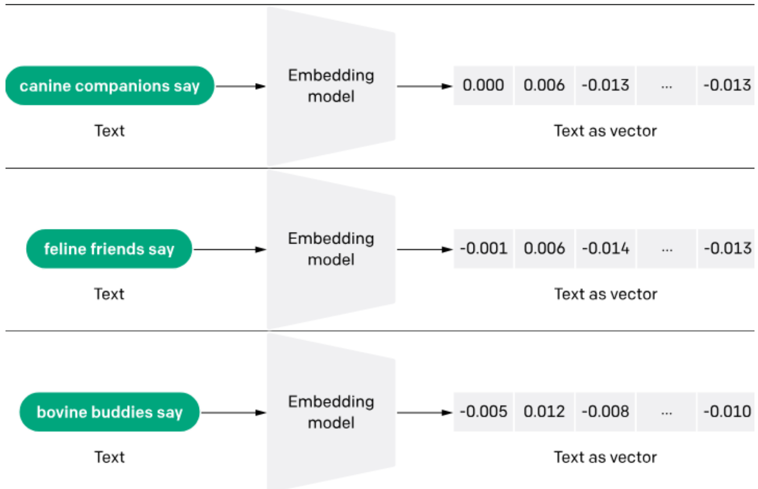

Figure 2.2: A basic overview of vectorisation (OpenAI, 2024a).

2.1.3 Large language models

LLMs are colossal machine learning models that leverage NLP to generate text, and have become widely used across industries in place of technical support and human resources (Vrontis et al., 2022). The training data required for an LLM is immense, reaching 45 terabytes of text data for ChatGPT in 2023 (Dwivedi et al., 2023).

This data is harvested from websites and social media due to them being the largest repositories of opinionated text data (Dubey et al. (2024), Z. Wang et al. (2016)). However, meticulous care is taken into the specific sources used to remove Personally Identifiable Information (PII) to minimise privacy and ethical concerns (Dubey et al., 2024).

The previously mentioned Transformer by Vaswani et al. (2017) became a staple in LLMs due to the major reduction in necessary processing power to produce higher-quality results, and it continues to underpin many LLMs today, including ChatGPT (Brown et al., 2020). Even with these enhancements, LLMs are still extremely performance intensive, requiring more than 8 top-range server-grade GPUs to run some of the most powerful high-parameter models like LLaMA 3.1's 405 billion parameter model (Dubey et al., 2024), and many therefore use cloud API solutions to access LLMs.

The amount of parameters in a model does not entirely account for the quality of its responses, as studied by Ouyang et al. (2022) in Figure 2.3 wherein their surveys revealed their fine-tuned LLM "InstructGPT" with over 100x less parameters than a 175 billion parameter GPT3 model would often give answers preferred by its human assessors, which reveals that the fine-tuning and prompt engineering of an LLM is as vitally important to the quality of its responses as the amount of parameters.

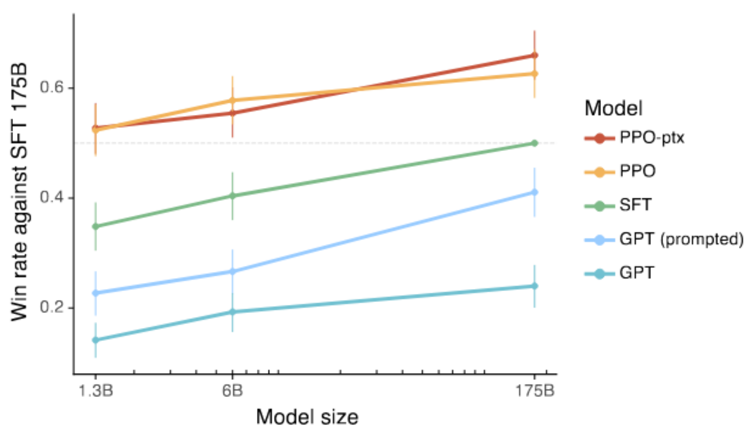

Figure 2.3: Human evaluations of the GPT models produced by Ouyang et al. (2022). PPO and PPO-ptx are their models.

The simplest way to measure the accuracy and quality of an LLM's responses is through human evaluation surveys such as that conducted by Ouyang et al. (2022), though software approaches such as DeepEval can be used. DeepEval offers 14 metrics to test LLM outputs with (DeepEval, 2024), with a notable metric being "G-Eval", originally introduced by Liu et al. (2023), which uses an "LLM-as-a-judge" approach where an LLM will evaluate and grade the quality of the output.

2.1.4 Retrieval-Augmented Generation

While LLMs are highly useful tools across many industries, they are not without limitations. The most notable of these limitations are hallucinations (P. Lewis et al., 2021), where the LLM will fabricate information that conflicts with user input, earlier conversation context or true facts (Zhang et al., 2023). This occurs as a direct result of the LLM's parametric memory[2] being overfitted or biased, which can be counteracted through introducing an external knowledge source, known as non-parametric memory (Komeili, Shuster and Weston (2022), Siriwardhana et al. (2023)).

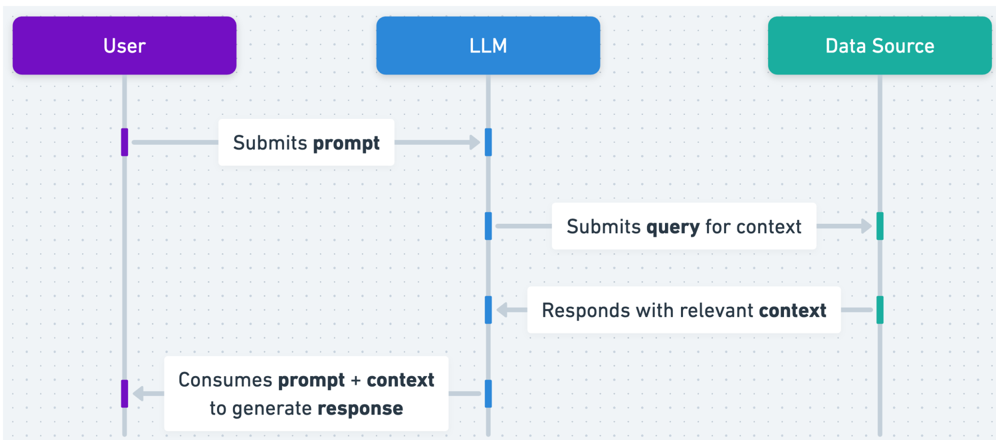

Figure 2.4: A basic overview of a RAG workflow (OpenAI, 2024b).

Siriwardhana et al. (2023) expanded upon the earlier works of Karpukhin et al. (2020) and M. Lewis et al. (2020) by creating "RAG-end2end", which explored the capabilities of RAG on a dynamically updating knowledge store, meaning the LLM itself would not have to be retrained every time the data updates, saving enormous amounts of processing power.

RAG is dependent upon external knowledge stores such as vector databases, which store and process vectorised data for non-parametric memory (Li, 2023), which makes them an essential part of the backend of a RAG-enabled chatbot as studied by Odede and Frommholz (2024).

Many software options exist for vector databases, such as Milvus (J. Wang et al., 2021), Pinecone (Pinecone, 2024) and Chroma (Chroma, 2024). Xie et al. (2023) compared these three, citing Pinecone's 'robust distributed computing capabilities and scalability', and its common usage in real-time searching scenarios.

Pinecone was also used in chatbots by Odede and Frommholz (2024) and Singer et al. (2024), showcasing its potential as a vector database solution for chatbots.

LangChain (LangChain, 2024) is a popular open-source framework for RAG pipelines that can be used to connect backend elements together, as described by Singer et al. (2024) when they used it to chunk their text data and connect to their Pinecone vector database to store the embedded data.

[2]Knowledge that the LLM has from its training data (Siriwardhana et al., 2023).

2.1.5 Chatbots / Conversational Agents

Conversational agents, better known as chatbots, leverage NLP in order to simulate a conversational flow between a user and machine, and have become mainstream products in recent years (Liao et al., 2018), though have existed as far back as 1966 with the creation of "ELIZA" for the IBM-7094 (Weizenbaum, 1966). As time has passed, advancements in chatbots have occurred in "waves", where each new wave has brought a major innovation (Schöbel et al., 2024).

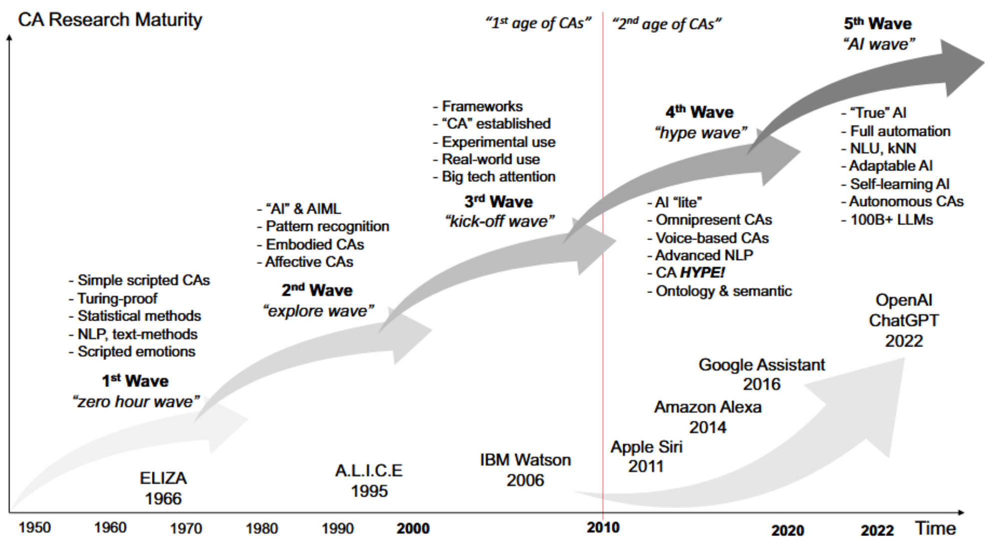

Figure 2.5: The five waves of conversational agent research (Schöbel et al., 2024).

Due to these considerable developments in the field, chatbots are now widely used across industries such as education (Kuhail et al., 2023). However, the use of the latest wave of chatbots based on LLMs poses significant risks, especially in educational settings as studied by Neumann et al. (2024), due to the risk of hallucinations being interpreted as absolute fact, although Shuster et al. (2021) argued that this risk can be greatly reduced through introducing RAG to the backend LLM, which is further backed by the RAG-based chatbot created by Ge et al. (2023), which they found to also give superior answers to those of a general-purpose chatbot without RAG.

Many platforms exist to aid chatbot development, though they are typically aimed at users from non-IT backgrounds (Srivastava and Prabhakar, 2020). Popular platforms include IBM's watsonx Assistant (IBM, 2024a), Google's Dialogflow (Google, 2024) and Microsoft's Bot Framework (Microsoft, 2024). However, these are primarily targeted at enterprise clients which is reflected in their pricing. Instead of using these, the chatbot can be manually developed using LangChain as its framework.

2.1.6 User experience and Human-Computer Interaction

The way people interact with their devices has drastically evolved over the years, from early MS-DOS command-line interfaces (CLIs) to mouse-based graphical user interfaces (GUIs), to touch screens (Kotian et al., 2024), greatly broadening the userbase of computers worldwide. Therefore, inclusive and accessible design is increasingly important to maximise the audience of any software, especially considering the growing disabled population (Putnam et al., 2012).

As well as being inclusive, the design should also be user-centred, meaning it should be an iterative process that is constantly taking user feedback into account (Chammas, Quaresma and Mont'Alvão, 2015). However, there are some barriers in this process when developing chatbots, as studied by Clark et al. (2019) in their survey of university students who stated that they view chatbots as tools, and would not converse with them in the same way as they would a person, which would limit their potential use and hinder the overall design process.

Users also often struggle to get chatbots to respond how they want, as their prompts may be poorly understood due to issues like overgeneralisation (Zamfirescu-Pereira et al., 2023), and studies show that they grow impatient after around 2 to 6 failed attempts, often branding the product as poor if this occurs (Luger and Sellen, 2016).

2.2 Summary

In conclusion, this literature review has revealed multiple key focus areas for the chatbot's development. The overall design of the chatbot must be iterative and human-centred, and user feedback should be obtained at every possible opportunity to ensure the resultant product is high quality.

A deep exploration into AI, specifically in its applications in NLP, LLMs and RAG, has revealed that the best approach will be to leverage a pre-existing cloud-based LLM, such as GPT-4o-mini, via an API, as running an LLM on a local machine would require an infeasible amount of processing power.

The non-parametric memory accessed through RAG would be a vector database created with Pinecone storing embeddings generated by OpenAI's text-embeddings-3-small model, and the overall framework will be LangChain. This will keep the cost of the project low while maintaining a tolerable level of quality in the bot's responses.

Methods and Implementation

This chapter focuses on the experimental design and implementation of the artefact.

3.1 Methodology

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Curabitur et erat consectetur, scelerisque eros nec, ultricies est. Donec mi ipsum, imperdiet vitae arcu quis, luctus venenatis quam. Integer ac massa a augue venenatis fringilla. Etiam posuere libero sed nulla tristique volutpat.

3.2 Design

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Curabitur et erat consectetur, scelerisque eros nec, ultricies est. Donec mi ipsum, imperdiet vitae arcu quis, luctus venenatis quam. Integer ac massa a augue venenatis fringilla. Etiam posuere libero sed nulla tristique volutpat.

3.3 Implementation

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Curabitur et erat consectetur, scelerisque eros nec, ultricies est. Donec mi ipsum, imperdiet vitae arcu quis, luctus venenatis quam. Integer ac massa a augue venenatis fringilla. Etiam posuere libero sed nulla tristique volutpat.

Evaluation

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Curabitur et erat consectetur, scelerisque eros nec, ultricies est. Donec mi ipsum, imperdiet vitae arcu quis, luctus venenatis quam. Integer ac massa a augue venenatis fringilla. Etiam posuere libero sed nulla tristique volutpat.

4.1 Methodology

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Curabitur et erat consectetur, scelerisque eros nec, ultricies est. Donec mi ipsum, imperdiet vitae arcu quis, luctus venenatis quam. Integer ac massa a augue venenatis fringilla. Etiam posuere libero sed nulla tristique volutpat.

4.1.1 Metrics

!!!

4.1.2 Baseline systems

4.1.3 Dataset

Likely not applicable. OpenAI's models are all closed-source.

4.2 Results

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Curabitur et erat consectetur, scelerisque eros nec, ultricies est. Donec mi ipsum, imperdiet vitae arcu quis, luctus venenatis quam. Integer ac massa a augue venenatis fringilla. Etiam posuere libero sed nulla tristique volutpat.

4.3 Discussion

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Curabitur et erat consectetur, scelerisque eros nec, ultricies est. Donec mi ipsum, imperdiet vitae arcu quis, luctus venenatis quam. Integer ac massa a augue venenatis fringilla. Etiam posuere libero sed nulla tristique volutpat.

Conclusions

"You should not include any new information or discussion in this section." This section must also link to the project's objectives.

Recommendations for future work

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Curabitur et erat consectetur, scelerisque eros nec, ultricies est. Donec mi ipsum, imperdiet vitae arcu quis, luctus venenatis quam. Integer ac massa a augue venenatis fringilla. Etiam posuere libero sed nulla tristique volutpat.

References

Chammas, Adriana, Manuela Quaresma and Cláudia Mont'Alvão (1st Jan. 2015). "A Closer Look on the User Centred Design". In: *Procedia Manufacturing*. 6th International Conference on Applied Human Factors and Ergonomics (AHFE 2015) and the Affiliated Conferences, AHFE 2015 3, pp. 5397–5404. ISSN: 2351-9789. DOI: 10.1016/j.promfg.2015.07.656.

Chroma (2024). *Chroma*. URL: https://www.trychroma.com/ (visited on 24/11/2024).

Clark, Leigh, Nadia Pantidi, Orla Cooney, Philip Doyle, Diego Garaialde, Justin Edwards, Brendan Spillane, Emer Gilmartin, Christine Murad, Cosmin Munteanu, Vincent Wade and Benjamin R. Cowan (2nd May 2019). "What Makes a Good Conversation? Challenges in Designing Truly Conversational Agents". In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. CHI '19. New York, NY, USA: Association for Computing Machinery, pp. 1–12. ISBN: 978-1-4503-5970-2. DOI: 10.1145/3290605.3300705.

DeepEval (22nd Nov. 2024). *Introduction | DeepEval - The Open-Source LLM Evaluation Framework*. URL: https://docs.confident-ai.com/docs/metrics-introduction (visited on 24/11/2024).

Devlin, Jacob, Ming-Wei Chang, Kenton Lee and Kristina Toutanova (June 2019). "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)*. NAACL-HLT 2019. Ed. by Jill Burstein, Christy Doran and Thamar Solorio. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 4171–4186. DOI: 10.18653/v1/N19-1423.

Dubey, Abhimanyu et al. (15th Aug. 2024). *The Llama 3 Herd of Models*. DOI: 10.48550/arXiv.2407.21783. arXiv: 2407.21783.

Dwivedi, Y.K., N. Kshetri, L. Hughes, E.L. Slade, A. Jeyaraj, A.K. Kar, A.M. Baabdullah, A. Koohang, V. Raghavan, M. Ahuja, H. Albanna, M.A. Albashrawi, A.S. Al-Busaidi, J. Balakrishnan, Y. Barlette, S. Basu, I. Bose, L. Brooks, D. Buhalis, L. Carter, S. Chowdhury, T. Crick, S.W. Cunningham, G.H. Davies, R.M. Davison, R. Dé, D. Dennehy, Y. Duan, R. Dubey, R. Dwivedi, J.S. Edwards, C. Flavián, R. Gauld, V. Grover, M.-C. Hu, M. Janssen, P. Jones, I. Junglas, S. Khorana, S. Kraus, K.R. Larsen, P. Latreille, S. Laumer, F.T. Malik, A. Mardani, M. Mariani, S. Mithas, E. Mogaji, J.H. Nord, S. O'Connor, F. Okumus, M. Pagani, N. Pandey, S. Papagiannidis, I.O. Pappas, N. Pathak, J. Pries-Heje, R. Raman, N.P. Rana, S.-V. Rehm, S. Ribeiro-Navarrete, A. Richter, F. Rowe, S. Sarker, B.C. Stahl, M.K. Tiwari, W. van der Aalst, V. Venkatesh, G. Viglia, M. Wade, P. Walton, J. Wirtz and R. Wright (2023). ""So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy". In: *International Journal of Information Management* 71. DOI: 10.1016/j.ijinfomgt.2023.102642.

Ge, Jin, Steve Sun, Joseph Owens, Victor Galvez, Oksana Gologorskaya, Jennifer C Lai, Mark J Pletcher and Ki Lai (1st Nov. 2023). "Development of a Liver Disease-Specific Large Language Model Chat Interface using Retrieval Augmented Generation". In: *medRxiv*, p. 2023.11.10.23298364. DOI: 10.1101/2023.11.10.23298364.

Google (2024). *Conversational Agents and Dialogflow*. Google Cloud. URL: https://cloud.google.com/products/conversational-agents (visited on 24/11/2024).

IBM (3rd Apr. 2024a). *IBM watsonx Assistant Virtual Agent*. URL: https://www.ibm.com/products/watsonx-assistant (visited on 24/11/2024).

Karpukhin, Vladimir, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen and Wen-tau Yih (Nov. 2020). "Dense Passage Retrieval for Open-Domain Question Answering". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. EMNLP 2020. Ed. by Bonnie Webber, Trevor Cohn, Yulan He and Yang Liu. Online: Association for Computational Linguistics, pp. 6769–6781. DOI: 10.18653/v1/2020.emnlp-main.550.

Komeili, Mojtaba, Kurt Shuster and Jason Weston (May 2022). "Internet-Augmented Dialogue Generation". In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. ACL 2022. Ed. by Smaranda Muresan, Preslav Nakov and Aline Villavicencio. Dublin, Ireland: Association for Computational Linguistics, pp. 8460–8478. DOI: 10.18653/v1/2022.acl-long.579.

Kotian, Abhijith L, Reshna Nandipi, Ushag M, Usha Rani S, VARSHAUK and Veena G T (Jan. 2024). "A Systematic Review on Human and Computer Interaction". In: *2024 2nd International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT)*. 2024 2nd International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT), pp. 1214–1218. DOI: 10.1109/IDCIoT59759.2024.10467622.

Krizhevsky, Alex, Ilya Sutskever and Geoffrey E Hinton (2012). "ImageNet Classification with Deep Convolutional Neural Networks". In: *Advances in Neural Information Processing Systems*. Vol. 25. Curran Associates, Inc. DOI: 10.1145/3065386.

Kuhail, M.A., N. Alturki, S. Alramlawi and K. Alhejori (2023). "Interacting with educational chatbots: A systematic review". In: *Education and Information Technologies* 28 (1), pp. 973–1018. DOI: 10.1007/s10639-022-11177-3.

LangChain (2024). *Introduction | LangChain*. URL: https://python.langchain.com/docs/introduction/ (visited on 24/11/2024).

Le, Quoc V. and Tomas Mikolov (22nd May 2014). *Distributed Representations of Sentences and Documents*. DOI: 10.48550/arXiv.1405.4053. arXiv: 1405.4053.

Lewis, Mike, Marjan Ghazvininejad, Gargi Ghosh, Armen Aghajanyan, Sida Wang and Luke Zettlemoyer (26th June 2020). *Pre-training via Paraphrasing*. DOI: 10.48550/arXiv.2006.15020. arXiv: 2006.15020.

Lewis, Patrick, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel and Douwe Kiela (12th Apr. 2021). *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks*. DOI: 10.48550/arXiv.2005.11401. arXiv: 2005.11401.

Li, Feifei (1st Aug. 2023). "Modernization of Databases in the Cloud Era: Building Databases that Run Like Legos". In: *Proc. VLDB Endow.* 16 (12), pp. 4140–4151. ISSN: 2150-8097. DOI: 10.14778/3611540.3611639.

Liao, Q. Vera, Muhammed Mas-ud Hussain, Praveen Chandar, Matthew Davis, Yasaman Khazaeni, Marco Patricio Crasso, Dakuo Wang, Michael Muller, N. Sadat Shami and Werner Geyer (19th Apr. 2018). "All Work and No Play?" In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. CHI '18. New York, NY, USA: Association for Computing Machinery, pp. 1–13. ISBN: 978-1-4503-5620-6. DOI: 10.1145/3173574.3173577.

Liu, Yang, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu and Chenguang Zhu (23rd May 2023). *G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment*. DOI: 10.48550/arXiv.2303.16634. arXiv: 2303.16634.

Luger, Ewa and Abigail Sellen (7th May 2016). “"Like Having a Really Bad PA": The Gulf between User Expectation and Experience of Conversational Agents”. In: *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. CHI '16. New York, NY, USA: Association for Computing Machinery, pp. 5286–5297. ISBN: 978-1-4503-3362-7. DOI: 10.1145/2858036.2858288.

Maedche, Alexander, Christine Legner, Alexander Benlian, Benedikt Berger, Henner Gimpel, Thomas Hess, Oliver Hinz, Stefan Morana and Matthias Söllner (1st Aug. 2019). “AI-Based Digital Assistants”. In: *Business & Information Systems Engineering* 61 (4), pp. 535–544. ISSN: 1867-0202. DOI: 10.1007/s12599-019-00600-8.

Microsoft (2024). *Microsoft Bot Framework*. URL: https://dev.botframework.com/ (visited on 24/11/2024).

Mikolov, Tomas, Kai Chen, Greg Corrado and Jeffrey Dean (7th Sept. 2013). *Efficient Estimation of Word Representations in Vector Space*. DOI: 10.48550/arXiv.1301.3781. arXiv: 1301.3781.

Miró-Nicolau, M., A. Jaume-i-Capó and G. Moyà-Alcover (2025). “A comprehensive study on fidelity metrics for XAI”. In: *Information Processing and Management* 62 (1). DOI: 10.1016/j.ipm.2024.103900.

Neumann, Alexander Tobias, Yue Yin, Sulayman Sowe, Stefan Decker and Matthias Jarke (2024). “An LLM-Driven Chatbot in Higher Education for Databases and Information Systems”. In: *IEEE Transactions on Education*. Conference Name: IEEE Transactions on Education, pp. 1–14. ISSN: 1557-9638. DOI: 10.1109/TE.2024.3467912.

Odede, Julius and Ingo Frommholz (10th Mar. 2024). “JayBot – Aiding University Students and Admission with an LLM-based Chatbot”. In: *Proceedings of the 2024 Conference on Human Information Interaction and Retrieval*. CHIIR '24. New York, NY, USA: Association for Computing Machinery, pp. 391–395. ISBN: 9798400704345. DOI: 10.1145/3627508.3638293.

OpenAI (2024a). *Introducing text and code embeddings*. URL: https://openai.com/index/introducing-text-and-code-embeddings/ (visited on 25/11/2024).

OpenAI (2024b). *Retrieval Augmented Generation (RAG) and Semantic Search for GPTs | OpenAI Help Center*. URL: https://help.openai.com/en/articles/8868588-retrieval-augmented-generation-rag-and-semantic-search-for-gpts (visited on 23/11/2024).

OpenAI (2024c). *Vector embeddings*. Vector embeddings. URL: https://platform.openai.com/docs/guides/embeddings (visited on 24/11/2024).

Ouyang, Long, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike and Ryan Lowe (4th Mar. 2022). *Training language models to follow instructions with human feedback*. DOI: 10.48550/arXiv.2203.02155. arXiv: 2203.02155.

Pinecone (2024). *Pinecone Documentation*. Pinecone Docs. URL: https://docs.pinecone.io/guides/get-started/overview (visited on 24/11/2024).

Putnam, Cynthia, Kathryn Wozniak, Mary Jo Zefeldt, Jinghui Cheng, Morgan Caputo and Carl Duffield (22nd Oct. 2012). "How do professionals who create computing technologies consider accessibility?" In: *Proceedings of the 14th international ACM SIGACCESS conference on Computers and accessibility*. ASSETS '12. New York, NY, USA: Association for Computing Machinery, pp. 87–94. ISBN: 978-1-4503-1321-6. DOI: 10.1145/2384916.2384932.

Samuel, A. L. (July 1959). "Some Studies in Machine Learning Using the Game of Checkers". In: *IBM Journal of Research and Development* 3 (3). Conference Name: IBM Journal of Research and Development, pp. 210–229. ISSN: 0018-8646. DOI: 10.1147/rd.33.0210.

Schöbel, Sofia, Anuschka Schmitt, Dennis Benner, Mohammed Saqr, Andreas Janson and Jan Marco Leimeister (1st Apr. 2024). "Charting the Evolution and Future of Conversational Agents: A Research Agenda Along Five Waves and New Frontiers". In: *Information Systems Frontiers* 26 (2), pp. 729–754. ISSN: 1572-9419. DOI: 10.1007/s10796-023-10375-9.

Selbst, Andrew D., Danah Boyd, Sorelle A. Friedler, Suresh Venkatasubramanian and Janet Vertesi (29th Jan. 2019). "Fairness and Abstraction in Sociotechnical Systems". In: *Proceedings of the Conference on Fairness, Accountability, and Transparency*. FAT* '19. New York, NY, USA: Association for Computing Machinery, pp. 59–68. ISBN: 978-1-4503-6125-5. DOI: 10.1145/3287560.3287598.

Shneiderman, Ben (16th Oct. 2020). "Bridging the Gap Between Ethics and Practice: Guidelines for Reliable, Safe, and Trustworthy Human-centered AI Systems". In: *ACM Trans. Interact. Intell. Syst.* 10 (4), 26:1–26:31. ISSN: 2160-6455. DOI: 10.1145/3419764.

Shuster, Kurt, Spencer Poff, Moya Chen, Douwe Kiela and Jason Weston (Nov. 2021). "Retrieval Augmentation Reduces Hallucination in Conversation". In: *Findings of the Association for Computational Linguistics: EMNLP 2021*. Findings 2021. Ed. by Marie-Francine Moens, Xuanjing Huang, Lucia Specia and Scott Wen-tau Yih. Punta Cana, Dominican Republic: Association for Computational Linguistics, pp. 3784–3803. DOI: 10.18653/v1/2021.findings-emnlp.320.

Singer, Maxwell B., Julia J. Fu, Jessica Chow and Christopher C. Teng (1st Mar. 2024). "Development and Evaluation of Aeyeconsult: A Novel Ophthalmology Chatbot Leveraging Verified Textbook Knowledge and GPT-4". In: *Journal of Surgical Education* 81 (3), pp. 438–443. ISSN: 1931-7204. DOI: 10.1016/j.jsurg.2023.11.019.

Siriwardhana, Shamane, Rivindu Weerasekera, Elliott Wen, Tharindu Kaluarachchi, Rajib Rana and Suranga Nanayakkara (12th Jan. 2023). "Improving the Domain Adaptation of Retrieval Augmented Generation (RAG) Models for Open Domain Question Answering". In: *Transactions of the Association for Computational Linguistics* 11, pp. 1–17. ISSN: 2307-387X. DOI: 10.1162/tacl_a_00530.

Srivastava, Saurabh and T.V. Prabhakar (Sept. 2020). "Desirable Features of a Chatbot-building Platform". In: *2020 IEEE International Conference on Humanized Computing and Communication with Artificial Intelligence (HCCAI)*. 2020 IEEE International Conference on Humanized Computing and Communication with Artificial Intelligence (HCCAI), pp. 61–64. DOI: 10.1109/HCCAI49649.2020.00016.

Turing, A. M. (1st Oct. 1950). "I.—COMPUTING MACHINERY AND INTELLIGENCE". In: *Mind* LIX (236), pp. 433–460. ISSN: 1460-2113, 0026-4423. DOI: 10.1093/mind/LIX.236.433.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser and Illia Polosukhin (4th Dec. 2017). "Attention is all you need". In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. NIPS'17. Red Hook, NY, USA: Curran Associates Inc., pp. 6000–6010. ISBN: 978-1-5108-6096-4. DOI: 10.48550/arXiv.1706.03762.

Vrontis, Demetris, Michael Christofi, Vijay Pereira, Shlomo Tarba, Anna Makrides and Eleni Trichina (26th Mar. 2022). "Artificial intelligence, robotics, advanced technologies and human resource management: a systematic review". In: *The International Journal of Human Resource Management* 33 (6). Publisher: Routledge, pp. 1237–1266. ISSN: 0958-5192. DOI: 10.1080/09585192.2020.1871398.

Wang, Jianguo, Xiaomeng Yi, Rentong Guo, Hai Jin, Peng Xu, Shengjun Li, Xiangyu Wang, Xiangzhou Guo, Chengming Li, Xiaohai Xu, Kun Yu, Yuxing Yuan, Yinghao Zou, Jiquan Long, Yudong Cai, Zhenxiang Li, Zhifeng Zhang, Yihua Mo, Jun Gu, Ruiyi Jiang, Yi Wei and Charles Xie (18th June 2021). "Milvus: A Purpose-Built Vector Data Management System". In: *Proceedings of the 2021 International Conference on Management of Data*. SIGMOD '21. New York, NY, USA: Association for Computing Machinery, pp. 2614–2627. ISBN: 978-1-4503-8343-1. DOI: 10.1145/3448016.3457550.

Wang, Zhaoxia, Chee Seng Chong, Landy Lan, Yinping Yang, Seng Beng Ho and Joo Chuan Tong (Dec. 2016). "Fine-grained sentiment analysis of social media with emotion sensing". In: *2016 Future Technologies Conference (FTC)*. 2016 Future Technologies Conference (FTC), pp. 1361–1364. DOI: 10.1109/FTC.2016.7821783.

Weizenbaum, Joseph (1st Jan. 1966). "ELIZA—a computer program for the study of natural language communication between man and machine". In: *Commun. ACM* 9 (1), pp. 36–45. ISSN: 0001-0782. DOI: 10.1145/365153.365168.

Wirtz, J., P.G. Patterson, W.H. Kunz, T. Gruber, V.N. Lu, S. Paluch and A. Martins (2018). "Brave new world: service robots in the frontline". In: *Journal of Service Management* 29 (5), pp. 907–931. DOI: 10.1108/JOSM-04-2018-0119.

Xie, Xingrui, Han Liu, Wenzhe Hou and Hongbin Huang (Dec. 2023). "A Brief Survey of Vector Databases". In: *2023 9th International Conference on Big Data and Information Analytics (BigDIA)*. 2023 9th International Conference on Big Data and Information Analytics (BigDIA). ISSN: 2771-6902, pp. 364–371. DOI: 10.1109/BigDIA60676.2023.10429609.

Zamfirescu-Pereira, J.D., Richmond Y. Wong, Bjoern Hartmann and Qian Yang (19th Apr. 2023). "Why Johnny Can't Prompt: How Non-AI Experts Try (and Fail) to Design LLM Prompts". In: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. CHI '23. New York, NY, USA: Association for Computing Machinery, pp. 1–21. ISBN: 978-1-4503-9421-5. DOI: 10.1145/3544548.3581388.

Zhang, Yue, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi and Shuming Shi (24th Sept. 2023). *Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models*. DOI: 10.48550/arXiv.2309.01219. arXiv: 2309.01219.

Bibliography

AWS (2024). *What is RAG? - Retrieval-Augmented Generation AI Explained - AWS*. Amazon Web Services, Inc. URL: https://aws.amazon.com/what-is/retrieval-augmented-generation/ (visited on 23/11/2024).

Cambridge Dictionary (2024). *Meaning of user experience in English*. URL: https://dictionary.cambridge.org/dictionary/english/user-experience (visited on 28/10/2024).

Cloudflare (2024). *What is a large language model (LLM)?* URL: https://www.cloudflare.com/en-gb/learning/ai/what-is-large-language-model/ (visited on 28/10/2024).

Confident AI (2024). *LLM Evaluation Metrics: The Ultimate LLM Evaluation Guide - Confident AI*. URL: https://www.confident-ai.com/blog/llm-evaluation-metrics-everything-you-need-for-llm-evaluation (visited on 24/11/2024).

Databricks (18th Oct. 2023). *Retrieval Augmented Generation*. Databricks. URL: https://www.databricks.com/glossary/retrieval-augmented-generation-rag (visited on 23/11/2024).

Elastic (2024). *What is Semantic Search? | A Comprehensive Semantic Search Guide*. URL: https://www.elastic.co/what-is/semantic-search (visited on 24/11/2024).

IBM (16th Aug. 2024b). *What is AI?* URL: https://www.ibm.com/topics/artificial-intelligence (visited on 28/10/2024).

IBM (11th Aug. 2024c). *What Is NLP (Natural Language Processing)? | IBM*. URL: https://www.ibm.com/topics/natural-language-processing (visited on 04/11/2024).

IBM (2024d). *What is generative AI?* URL: https://research.ibm.com/blog/what-is-generative-AI (visited on 28/10/2024).

ICO (2024). *Definitions*. URL: https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/artificial-intelligence/explaining-decisions-made-with-artificial-intelligence/part-1-the-basics-of-explaining-ai/definitions/ (visited on 28/10/2024).

MIT (9th Nov. 2023). *Explained: Generative AI*. URL: https://news.mit.edu/2023/explained-generative-ai-1109 (visited on 28/10/2024).