Human vs. Machine: A Comparative Study on the Detection of AI-Generated Content

AMAL BOUTADJINE, Department of Computer Science, Université Ferhat Abbas, Setif, Algeria
FOUZI HARRAG, Computer Sciences, Universite Ferhat Abbas, Setif, Algeria
KHALED SHAALAN, Department of Informatics, British University - Faculty of Engineering and IT, Dubai, United Arab Emirates

The surge in advancements in large language models (LLMs) has expedited the generation of synthetic text imitating human writing styles. This, however, raises concerns about the potential misuse of synthetic textual data, which could compromise trust in online content. Against this backdrop, the present research aims to address the key challenges of detecting LLMs-generated texts. In this study, we used ChatGPT (v 3.5) because of its widespread and capability to comprehend and keep conversational context, allowing it to produce meaningful and contextually suitable responses. The problem revolves around the task of discerning between authentic and artificially generated textual content. To tackle this problem, we first created a dataset containing both real and DeepFake text. Subsequently, we employed transfer-learning (TL) and conducted DeepFake-detection utilizing SOTA large pre-trained LLMs. Furthermore, we conducted validation using benchmark datasets comprising unseen data samples to ensure that the model's performance reflects its ability to generalize to new data. Finally, we discussed this study's theoretical contributions, practical implications, limitations and potential avenues for future research, aiming to formulate strategies for identifying and detecting large-generative-models' produced texts. The results were promising, with accuracy ranging from 94% to 99%. The comparison between automatic detection and the human ability to detect DeepFake text revealed a significant gap in the human capacity for its identification, emphasizing an increasing need for sophisticated automated detectors. The investigation into AI-generated content detection holds central importance in the age of LLMs and technology convergence. This study is both timely and adds value to the ongoing discussion regarding the challenges associated with the pertinent theme of "DeepFake text detection", with a special focus on examining the boundaries of human detection.

CCS Concepts: • **Computing methodologies** → **Natural language processing**;

Additional Key Words and Phrases: Language processing, Large Language Models, Generative AI, AI-Generated Content detection, Comparative Study, ChatGPT

ACM Reference Format:
Amal Boutadjine, Fouzi Harrag, and Khaled Shaalan. 2025. Human vs. Machine: A Comparative Study on the Detection of AI-Generated Content. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* 24, 2, Article 12 (February 2025), 26 pages. https://doi.org/10.1145/3708889

Authors' Contact Information: Amal Boutadjine, Department of Computer Science, Université Ferhat Abbas, Setif, Sétif, Algeria; e-mail: boutadjine.amal@univ-setif.dz; Fouzi Harrag, Computer Sciences, Universite Ferhat Abbas, Setif, Setif, Algeria; e-mail: fouzi.harrag@gmail.com; Khaled Shaalan, Department of Informatics, British University—Faculty of Engineering and IT, Dubai, Dubai, United Arab Emirates; e-mail: khaled.shaalan@buid.ac.ae.

ACM 2375-4699/2025/02-ART12
https://doi.org/10.1145/3708889

1 Introduction and Motivation

1.1 Preliminary

In the rapidly evolving world of internet-mediated information dissemination, with the strides achieved in synthetic image generation, significant progress in **natural language generation** (**NLG**) has empowered **deep learning** (**DL**) models to produce synthetic text of remarkable linguistic quality. Consequently, the exponential proliferation of textual content produced daily and dispersed throughout the digital landscape has given rise to multifaceted challenges. The dramatic advancements witnessed in **natural language processing** (**NLP**) have been propelled by the advent of Transformers [52] and their ensuing variant language models such as GPT-3 [8], T5 [44], LLaMA [50], Bard[1], and ChatGPT[2]. These advancements have orchestrated a revolutionary transformation, encompassing a spectrum of applications spanning text completion [48], question answering [43, 56, 66, 67], data-to-text generation [62, 63], story generation [24], sentiment analysis [27, 64, 65], document classification [42], word sense disambiguation [68], and many others.

As the utilization of these **Large Language Models** (**LLMs**) has proliferated effortlessly across the internet, the advent of LLMs has functioned as a catalyst, precipitating substantial advancements in generating text that exhibits striking coherence, consistency, and convincing semblance to human-authored content [21]. This has engendered a situation where the demarcation between human-produced and deep-synthesized textual content on the internet assumes paramount importance [11]. Consequently, the research and development of robust DeepFake-text detection methods have garnered significant attention and have evolved into a prominent focal point of scholarly investigation in recent years, spurred by the imperative to facilitate efficacious AI-generated-text detectors across diverse domains [13, 31, 46].

The efficacy of text classifier models hinges on their discerning capacity to unravel intricate patterns and relationships woven within textual data, rendering them pivotal tools in navigating the data inundation characterizing today's digital milieu. In response, researchers have embarked on an exploration of a spectrum of techniques, from classical **machine learning** (**ML**) algorithms [1, 28, 53] to state-of-the-art DL models [4, 21, 26] with the overarching goal of achieving precise and accurate classification results. The complexities and challenges inherent in DeepFake-text detection, such as the management of imbalanced datasets, the accommodation of contextually dependent semantics, and the assimilation of multilingual content, continue to fuel innovation within this domain.

While a reasonable body of research has emerged to address the detection of DeepFake-text in the English language, a notable gap remains in other languages, especially those with fewer resources, such as Arabic. This underlines the importance of extending this line of inquiry to encompass a broader linguistic landscape. Consequently, the objective of this research article is to augment the existing compendium of knowledge by introducing a novel methodology tailored for Arabic DeepFake-text detection. Our intent is to harness insights from recent advancements in DL, concurrently accounting for the intricacies inherent in the domain-specific data under scrutiny. By articulating the nuances of our proposed approach and meticulously evaluating its performance, we anticipate contributing substantively to the ongoing conversation surrounding techniques and applications pertinent to DeepFake-text detection. Through this research, we also aspire to systematically scrutinize and comprehensively assess the discernment abilities of internet users in their identification of DeepFake-text within authentic, real-world scenarios.

This study investigates the efficacy of four LLMs in recognizing Arabic DeepFake text produced by ChatGPT (version 3.5) and analyses human proficiency in identifying text generated by

[1] "Bard: OpenAI's Conversational AI Service." Bard, Google AI, available at: https://bard.google.com
[2] "ChatGPT (Version 3.5)." Available at: https://chat.openai.com/chat

ChatGPT. The comprehensive viewpoint highlights the importance of the present study, filling gaps in the literature, contributing to the continuing dialogue, and enhancing our collective comprehension of the issue.

The primary objective of this study was to create and assess robust **DeepFake Text Detectors** (**DFTDs**) tailored for Arabic text by utilizing transfer learning from cutting-edge multilingual language models. Through a comprehensive examination of four distinct pre-trained models (mBERT and XLM-RoBERTa variants), we aimed to overcome the significant deficiency in Arabic DeepFake text detection while evaluating the efficacy of automated detection in comparison to human capabilities.

Specifically, our research objectives encompassed the following:

- —To construct a comprehensive, high-quality gold standard dataset of Arabic DeepFake text to serve as a benchmark for detection system development.
- —To design, develop, and evaluate a robust and generalizable detection system capable of identifying Arabic DeepFake text effectively.
- —To assess the ability of individuals to distinguish between authentic and DeepFake textual content in digital contexts, highlighting the reliability of their discernment.

With respect to the stated aim and objectives, we primarily addressed the following **research questions** (**RQs**):

RQ1: How effective are the proposed models in identifying Arabic DeepFake text using the developed gold standard dataset?
RQ2: How well do the proposed detection models generalize to unseen data from other benchmark datasets?
RQ3: How does the proposed detection system compare to the state-of-the-art models in terms of key metrics such as precision, recall, F1-score, and accuracy?
RQ4: How accurately can internet users distinguish between authentic and DeepFake textual content, and how reliable are their judgments?

1.2 Contributions of the Paper

Previous research primarily centered on the English language with only a limited number of studies involving other languages. The originality of our work and the key contributions are listed below:

- —The introduction of DeepFake Text Detectors: DFTD1 (DeepFake Text Detector 1), DFTD2 (DeepFake Text Detector 2), DFTD3 (DeepFake Text Detector 3), and DFTD4 (DeepFake Text Detector 4) which utilize transfer learning by fine-tuning four LLMs for the detection of ChatGPT DeepFake text.
- —Developing a novel Arabic dataset of authentic text and DeepFake text generated using ChatGPT.
- —Designing and conducting experiments to assess the performance of the proposed models in comparison to alternative detectors.
- —Conducting an analytic investigation of human proficiency in recognizing DeepFake text and evaluating the outcomes.

A novel approach for recognizing DeepFake text is created to achieve the research aims. We also generated a new dataset of authentic/inauthentic Arabic text. We trained several LLMs on our Real/Fake dataset for the detection task and evaluated their performance against human capabilities in identifying AI-generated text.

The remainder of this article is structured as follows: Section 2 expounds upon the related literature in the domain of DeepFake text detection. Section 3 delineates the methodological approach harnessed in this investigation while Section 4 represents the experiments conducted in this study with the results obtained and the discussions. Subsequently, Section 5 encompasses theoretical, practical, and social implications. Finally, concluding insights are drawn in Section 6, supplemented by an exploration of limitations and future lines of our research.

2 Related Works

2.1 AI-Generated Text Detection

The investigation into identifying text generated by LLMs has gained significant traction within the academic and media landscape, reflecting the pressing need to discern between human-authored content and machine-generated text in order to avoid misleading audience and various assessors.

Chaka [12] evaluated five AI-tools for detecting AI-generated content, including Copyleaks AI Content Detector and GPT-2 output detector, with Copyleaks AI Content Detector demonstrating superior accuracy in recognizing AI-generated responses from ChatGPT, YouChat, and Chatsonic. According to Chaka [12], all five AI content detectors seem to have the same drawback of being unable to accurately and convincingly identify AI-generated writings in various settings. Alamleh et al. [1] conducted a study that gathered responses from computer science students regarding essay and programming assignments to evaluate the efficacy of various ML algorithms, such as **Support Vector Machines** (**SVM**), **Logistic Regression** (**LR**), and **Decision Trees** (**DT**), in distinguishing between human-written and AI-generated text. Their work explored applications in content moderation and plagiarism detection, achieving high accuracy in distinguishing the two text sources. The study's stated results indicated that traditional ML methods (Random Forest and Extremely Randomized Trees) outperformed the neural classifier (LSTM).

Wu et al. [58] presented LLMDet, a tool that employs perplexity scores and self-watermarking to ascertain whether text originates from a LLM or is authored by a human. Given that the perplexity calculation necessitates transparent access to token-level log probabilities, which is unfeasible in practical applications, the authors suggested calculating a proxy perplexity for each target LLM utilizing standard n-gram probabilities. These probabilities serve as the LLM's writing signature to identify the closest source to the input text's proxy perplexity.

Kumarage et al. [32] employed GPT-2 to produce text and compare it with content provided by humans. The suggested approach utilized stylometric signals, such as lexical and syntactic features, to improve the detection of AI-generated content by analysing the writing style of the text and recognizing distinctive patterns and characteristics. Upon extraction from the input text, these signals facilitate the sequence-based identification of tweets generated by AI. Their research underscored the efficacy of stylometric analysis in combating the proliferation of machine-generated content. Sadasivan et al. demonstrated in their study [46] the vulnerabilities of AI text detectors to paraphrase attacks, highlighting the necessity for resilient detection methods capable of withstanding such evasion strategies.

Antoun et al. [4] created ChatGPT detectors tailored for French text by translating existing English datasets and training classifiers. The authors demonstrated the challenges faced by advanced classifiers, trained on a blend of text generated by LLMs and human material, in identifying hostile literature composed in an academic, pedagogical, or encyclopedic manner. Weber-Wulff et al. [57] performed a comparative examination of multiple tools for identifying AI-generated content, emphasizing the shortcomings and constraints of existing technology in this domain. Fagni et al. [21] presented the TweepFake dataset, highlighting the difficulties in detecting DeepFake tweets.

Table 1. Comparative Characteristics of Related Work with the Present Study

Study	Detected	Language	Datasets	Detection Method
[1]	ChatGPT-generated text	English	Kaggle dataset	TF-IDF + ML algorithms
[32]	AI-generated tweets	English	• in-house dataset • TweepFake	stylometric signals
[4]	ChatGPT-generated text	English French	Translated dataset	Fine-tuning (RoBERTa, ELECTRA, CamemBERT, CamemBERTa)
[71]	LLMs-generated text	English	Open-source dataset	LSTM, Transformer and CNN
[72]	LLMs-generated text	English	DAIGT	Adaptive Ensembles of Fine-Tuned Transformers
[28]	GenAI-generated text	English	Written + generated passages	Human detection
[30]	ChatGPT-generated text	English	• human dataset • ChatGPT dataset	Tunicate Swarm Algorithm+ LSTMRNN
Current Study	LLMs-generated text	Arabic	• Ara-Deep • M4 • LLM Question-Answer Dataset • BLOOM Dataset	Fine-tuning DF1, DF2, DF3 and DF4

Recent investigations have concentrated on the detection of text created by ChatGPT across several domains. Katib et al. [30] introduced the TSA-LSTMRNN model to differentiate between ChatGPT-generated text and human writing. Elkhatat et al. [20] examined the effectiveness of AI content detectors in recognizing text created by ChatGPT on engineering subjects. Zhenyu et al. [61] developed a method for detecting ChatGPT-generated code using targeted masking perturbation, fulfilling the demand for dependable detection techniques in programming. Liu et al. [69] approached AI text detection as an authorship attribution problem, employing a strided sliding window method based on GPT-2 to extract perplexity features for differentiating text kinds. The magnitudes of perplexity features were evaluated to distinguish between AI-generated and human-generated text.

Kumarage et al. [70] conducted an evaluation of forensic systems for AI-generated text. It presented a taxonomy centered on detection, attribution, and characterization, demonstrating that detection Challenges encompass sustaining accuracy in the face of advancing AI technologies, necessitating future enhancements for the adaptability of forensic systems.

Mo et al. [71] proposed a model that integrates LSTM, Transformer, and CNN layers. The bidirectional LSTM was employed for sequence feature extraction, the multi-head attention mechanism to identify global dependencies, and the one-dimensional convolutional layer for local feature extraction, serving as a model for future study in AI text identification.

Notwithstanding considerable advancements in AI-generated text detection, the domain continues to encounter substantial obstacles. The majority of research has concentrated on English and other well-resourced languages, resulting in a deficiency in Arabic text identification. Furthermore, the swift advancement of AI models presents continuous hurdles to detection techniques. This study mitigates these constraints by creating advanced detectors tailored for Arabic machine-generated text. This research significantly contributes to the field by developing innovative DeepFake Text Detectors utilizing fine-tuned LLMs, establishing a new Arabic dataset, and evaluating model performance against human detection capabilities, thereby offering crucial tools and insights for addressing the proliferation of DeepFake content in a vital linguistic context. Table 1 reports comparative characteristics of related work with the present study.

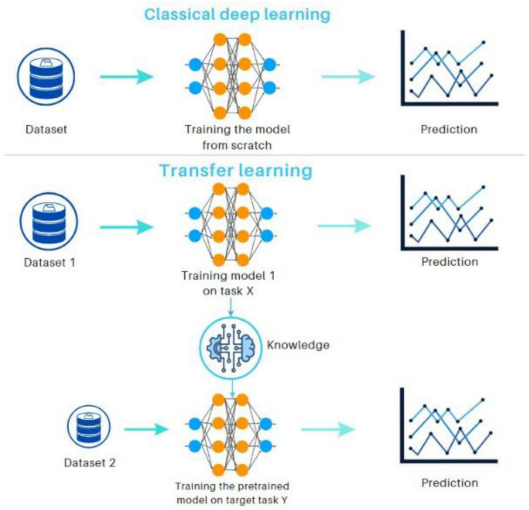

Fig. 1. Classical machine learning vs transfer learning process.

2.2 Transfer Learning

Transfer learning is a crucial paradigm that enables the extraction of knowledge from one domain or task and its application to another, enhancing model performance. This knowledge migration strategy offers an advantage by mitigating the scarcity of labeled data in target domains and concurrently expediting the learning curve through the utilization of shared underlying patterns.

Pan and Yang [41], define transfer learning as the enhancement of target domain learning through the integration of knowledge gained from a source domain. This knowledge transfer includes a wide range of information, encompassing feature representations, parameters, or even entire models, thus fostering heightened adaptability and generalization.

Transfer learning is widely adopted in recent NLP research [2, 5, 45], aided by the enormous development of LLMs. In NLP, transfer learning holds substantial significance for a variety of tasks, such as machine translation, fact-checking [38, 25], anomaly detection [39], text classification [23, 40, 60], and multitask learning in NLP [15].

The rise of transformer models driven by Vaswani et al. [52] breakthrough ''Attention is All You Need'' laid the groundwork for advancements like BERT. BERT, a deep bidirectional transformer for language understanding, was developed by [17], with additional refinements by Liu et al. [35] with RoBERTa and [47] with DistilBERT, exhibiting the trajectory of transfer learning's progress in NLP. BERT and its variations are commonly utilized in transfer learning [18, 54, 37]. Furthermore, other pre-trained language models [44, 10] have demonstrated excellent performance on NLP tasks.

As shown in Figure 1, traditional ML starts learning from scratch on a large dataset of labelled data. This procedure is considered highly time-consuming and computationally expensive, especially if the dataset is large or complex. On the contrast, transfer learning starts with a model that has already been trained on a sizable dataset containing labelled (or unlabeled) data. Therefore, the learning process is remarkably accelerated, allowing to train models using much smaller datasets and less processing power.

The research findings encompass both traditional ML and advanced DL systems, as indicated by the literature review of related research. The research findings indicate a continual difficulty in reliably identifying human-written texts (exceeding 80%), yet detecting AI-generated sections

continues to pose challenges for existing detection methods. Tools may demonstrate bias, particularly when distinguishing between human-authored and ChatGPT-generated content, especially in instances of deliberate rewriting or paraphrasing. This bias substantially reduces detection efficacy. Prior research highlights that detecting ChatGPT-generated content necessitates both a suitable algorithm and a thorough comprehension of linguistic complexities.

3 Methodology

This study employs transfer learning on extensive pre-trained models to establish a method for detecting Arabic DeepFake text. Our methodology consists of three primary phases: the initial phase involves dataset generation, the second phase entails DeepFake text detection utilizing the proposed classifiers, and the third phase, detailed in Section 4.6, explores a comprehensive comparison between automated detection and human discernment of DeepFake text. We developed a novel dataset, Ara-Deep, utilizing the ChatGPT model (version 3.5) for the training data. our dataset was utilized to train our four models, DFTD1, DFTD2, DFTD3, and DFTD4, which were subsequently assessed on alternative datasets, establishing our study as one of the initial pioneering efforts in this domain to tackle the identification of Arabic DeepFake text generated using ChatGPT. Sections 3.2 and 3.3 illustrate the dataset generation methodology and the comprehensive architecture of the proposed models.

3.1 Problem Definition

DeepFake text detection is a challenging problem from NLP and AI perspectives. Similar to the concept of detecting DeepFake videos in the computer vision domain, this problem revolves around the task of discerning between authentic textual content and artificially generated text created by advanced LLMs. Given the increased skill of language models at producing grammatically correct and contextually coherent passages, the main difficulty lies in creating effective approaches to discriminate between authentic and synthesized textual content.

The problem can be framed as a binary classification task (with two distinguished classes: real text and DeepFake text). Each input text sample x_i is assigned a label $y_i \in \{0, 1\}$, with 0 denoting authentic text and 1 representing DeepFake text. The goal is to train the detector to learn a function: $f : x_i \mapsto y_i$ that generalizes well to previously unseen text instances, while being resilient to the evolving sophistication of language models utilized in generating synthetic content. In this study, the text authored by a human is referred to as "real text".

3.2 Dataset

Our custom dataset, so-called Ara-Deep, represents a fusion of artificially generated and authentic text data. The process of generating artificial text involves a meticulous prompt engineering approach, wherein real examples from the SANAD dataset were employed [19]. These real examples served as input to ChatGPT (version 3.5), guiding the model in rephrasing and completing text to produce coherent, natural, and human-like DeepFake text. Refer to Appendix A for additional information regarding the dataset creation and prompt engineering procedure.

As outlined in Table 2, we provide insights into our generated dataset's composition in terms of sentence count, word count, and unique words count.

Figure 2(a) provides a visual representation of the word frequency distribution within our dataset, while Figure 2(b) extends this representation to line distribution. For a more lucid global visualization, the dataset is split into ten distinct and equal segments in these renderings.

As shown in Figure 4, during the prompt engineering phase, text reformulation and text completion were employed, with the text being entered either at the beginning or the end of the prompt.

Table 2. Characteristics and Statistical Information of the Produced Dataset

Property	Value
#sentences	19,807
#words	2,065,670
#Unique_words	55,153
Average_words_per_sentence	105/36

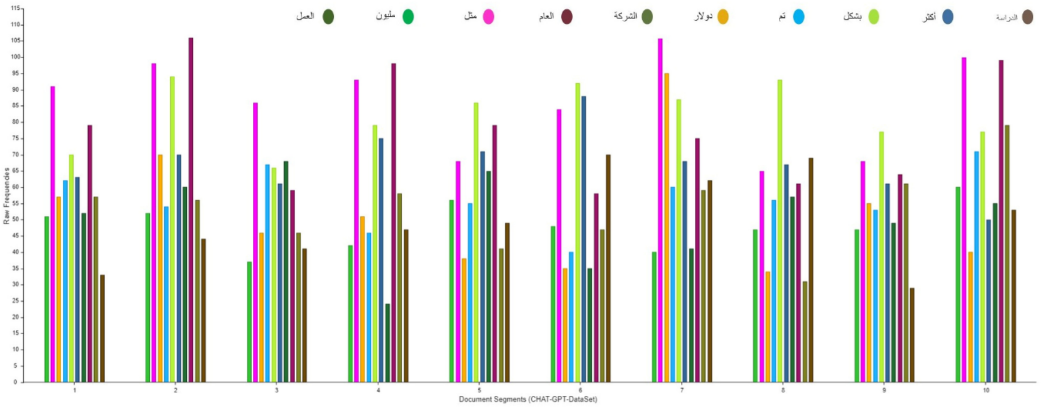

Fig. 2. Frequencies distribution of the top 10 words in our dataset. The whole dataset is segmented into10 segments.

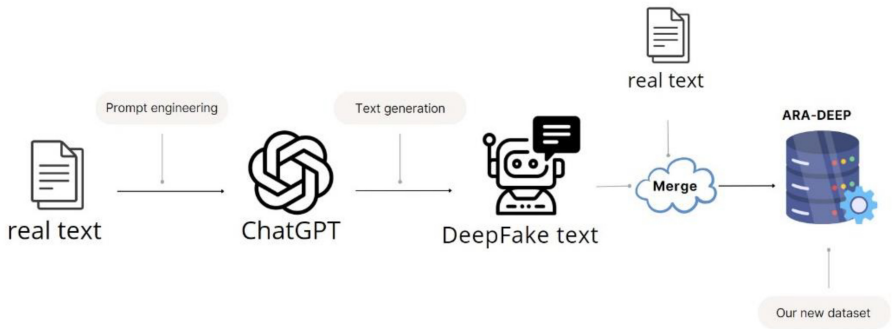

Fig. 3. The dataset construction pipeline.

The dataset generation pipeline is visually expounded upon in Figure 3. Table 3 offers an overview of the real text that was given as input and the corresponding DeepFake generated text samples.

To improve data quality before using the dataset for the training, we clean the data by removing URLs, hashtags, and symbols. The process was performed using Python with Pandas libraries, yielding a dataset containing 2,065,670 terms and 19,807 phrases, with a total of 55,153 unique words among these terms.

A word cloud representation of the most meaningful and visible words included in the Ara-Deep dataset can be seen in Figure 5.

Table 3. Two Examples of the Instances Included in Our Dataset Generated by Chatgpt

	Input text		Output text (DeepFake text)
	Prompt	real text	
text completion	الآتي النص أكمل:	الاقتصادية المدينة إعمار شركة أرباح شهدت	المالية للتقارير وفقًا .الماضية الفترة خلال قويًا نموًا الاقتصادية المدينة إعمار شركة أرباح شهدت ..ملحوظ بشكل أرباحها ارتفعت حيث ممتازًا، أداءها الشركة سجلت الأخيرة، .المتنوعة العقارية مشاريعها وإدارة تنمية في الشركة استراتيجية لنجاح إلى الأرباح في الزيادة هذه تعزى مختلفة قطاعات في والمبتكرة الكبيرة المشاريع من مجموعة بتطوير الاقتصادية المدينة إعمار قامت حيث .والترفيه والضيافة والتجارية السكنية العقارات مثل العام والاستقرار المحلي الاقتصاد قوة الاقتصادية المدينة إعمار لشركة الإيجابية الأرباح نتائج تعكس المشاريع إلى الاستثمارات من المزيد وتجذب المستثمرين بهذه النتائج هذه تعزز كما .العقاري للسوق .المدينة في العقارية رؤية إطار في والتكنولوجيا الابتكار وتعزيز الدخل مصادر تنويع نحو الاقتصادية المدينة إعمار توجهت ومؤسسات مع شركات والتجاري الاقتصادي التعاون لتعزيز الشركة تسعى كما .الأجل طويلة استراتيجية .والاستدامة النمو تحقيق بهدف دولية،	
	Complete the following text:	Emaar The Economic City's profits have witnessed	"Emaar The Economic City"'s profits have witnessed strong growth over the past period. According to recent financial reports, the company has recorded excellent performance, as its profits have increased significantly. This increase in profits is due to the success of the company's strategy in developing and managing its various real estate projects. Emaar The Economic City has developed a group of large and innovative projects in various sectors such as residential and commercial real estate, hospitality and entertainment. The positive earnings results of Emaar The Economic City reflect the strength of the local economy and the general stability of the real estate market. These results also enhance investor confidence and attract more investments to real estate projects in the city. Emaar The Economic City has moved towards diversifying sources of income and enhancing innovation and technology within the framework of a long-term strategic vision. The company also seeks to enhance economic and commercial cooperation with international companies and institutions, with the aim of achieving growth and sustainability. …
text reformulation	الآتي النص صياغة أعد:	الشاشات طي عملية تسهل ميكانيكية لآلية جديدة اختراع براءة حق على سامسونغ شركة حصلت بأنها تقدمت فقد كانت، 2017 عام خلال بالأسواق لها نموذج أول رؤية المتوقع من الذكية، الهواتف 2015. (حزيران) يونيو في أن شأنها من والتي الشاشة، خلف مثبتة ميكانيكية أجزاء عبر الهاتف شاشة طي على الآلية هذه وتعمل المفتوح الوضع في الشاشة تثبيت على تعمل كما البعض، بعضهما على الشاشة نصفي وغلق فتح عملية تدعم وأن سبق حيث سامسونغ، تسجلها التي نوعها من الأولى الاختراع براءة هذه تعد ولا .الاستخدام عند المزودة التقالب الهواتف صنع من تسهل وآليات لتقنيات اختراع براءات عدة الكورية الشركة سجلت وجدت وأنها لاسيما الفكرة، من الاقتصادية الجدوى مدى سامسونغ تبحث تزال ولا .القطبي قابلة بشاشات .الفكرة يجيدون لا المستهلكين بعض	طي شاشات عملية تسهل ميكانيكية لآلية جديدة اختراع براءة على الحصول من سامسونغ شركة تمكنت .2017 عام خلال الأسواق في التقنية هذه باستخدم نموذج أول نشهد أن من المتوقع .الذكية الهواتف مطويّ بشكل الشاشة وإغلاق فتح تمكين الشاشة خلف مثبتة ميكانيكية أجزاء على الآلية هذه تعتمد النوع، هذا في سامسونغ تُسجّلها اختراع براءة أول هي هذه وليست .الاستخدام عند استقرارها وتأمين ذات التقالب الهواتف صنع تسهّل وآليات لتقنيات اختراع براءات عدة الكورية الشركة قدّمت فقد خاصة الاقتصادية، الناحية من الفكرة هذه جدوى تدرس سامسونغ زالت وما .القطبي القابلة الشاشات .الفكرة هذه يفضلون لا الذين المستهلكين بعض وجود مع
	Reformulate the following text:	Samsung has obtained a new patent for a mechanism that facilitates the process of folding screens for smartphones. It is expected to see its first model on the market during 2017, which it filed in June 2015. This mechanism works to fold the phone screen through mechanical parts installed behind the screen, which will support the process of opening and closing the two halves of the screen to each other, and also works to hold the screen in the open position when in use. This is not the first patent of its kind registered by Samsung, as the Korean company has previously registered several patents for technologies and mechanisms that facilitate the manufacture of mobile phones equipped with foldable screens. Samsung is still examining the economic feasibility of the idea, especially since it found that some consumers do not like the idea.	Samsung was able to obtain a new patent for a mechanism that facilitates the process of folding smartphone screens. It is expected that we will see the first model using this technology on the market during 2017. This mechanism relies on mechanical parts installed behind the screen to enable the screen to be opened and closed in a folded manner and to ensure its stability when in use. This is not the first patent registered by Samsung in this type, as the Korean company has filed several patents for technologies and mechanisms that facilitate the manufacture of mobile phones with foldable screens. Samsung is still studying the economic feasibility of this idea, especially with some consumers who do not prefer this idea.

The first example is generated using text completion, while the second is produced based on a text paraphrasing prompt.

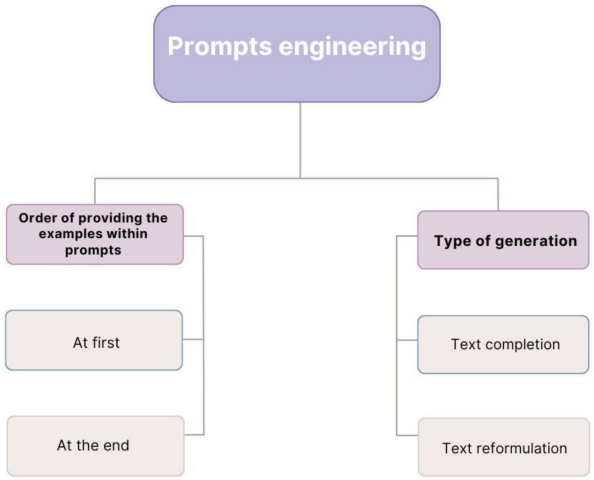

Fig. 4. Prompts engineering techniques used with ChatGPT for generating AI-produced text.

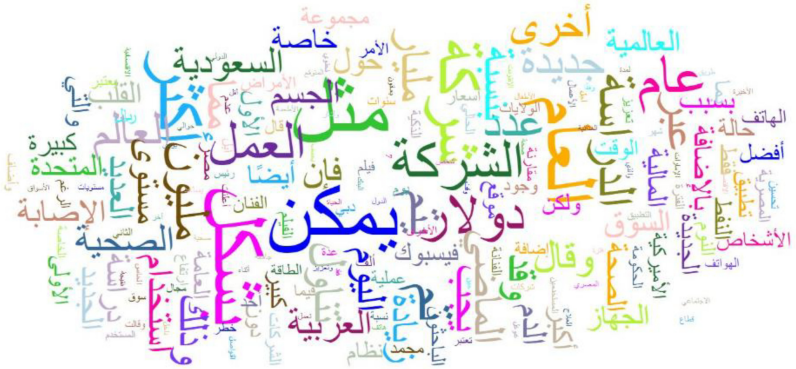

Fig. 5. Word cloud of the top words included in the dataset.

3.3 Deepfake Text Detection Models

As previously introduced, the current task of DeepFake text detection is approached as a binary classification problem. In our pursuit of delving into the ability to distinguish text generated by LLMs, our study conducted a meticulous investigation involving a cohort of four distinct pre-trained models, trained on multi-languages including Arabic and that performed well on NLU tasks. Notably, the models encompassed **Multilingual-BERT** (**mBERT**) [17], xlm-roberta-large, xlm-roberta-base, and xlm-roberta-large-xnli [16]. We judiciously selected these models for their established prominence and demonstrated capabilities in NLU and linguistic feature extraction [6, 34, 49, 59].

Our transfer learning approach leveraged these pre-trained models' rich linguistic knowledge acquired from their extensive pre-training on multilingual corpora. Specifically:

—DFTD1 transferred knowledge from fine-tuning mBERT, an embodiment of the BERT architecture, which has pretrained deep bidirectional representations from unlabeled text by

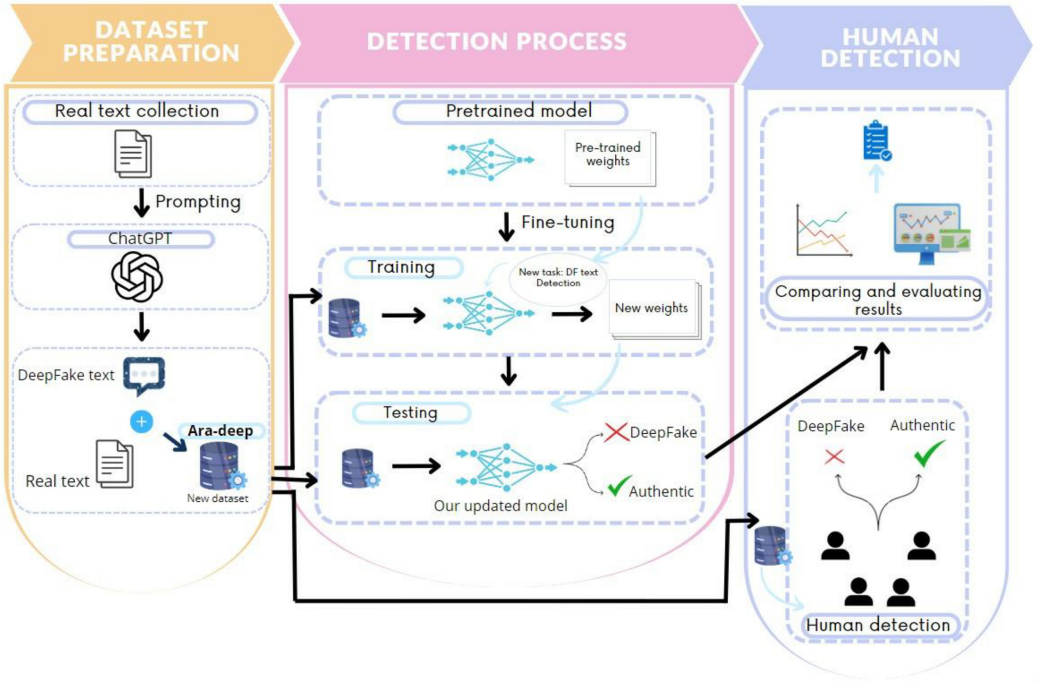

Fig. 6. The overall architecture of the proposed model.

conditioning on both left and right context in all layers and serves as a universal transformer model that exhibits proficiency in various languages.

— DFTD2, DFTD3, and DFTD4 transferred knowledge from fine-tuning xlm-roberta-base, xlm-roberta-large, and xlm-roberta-large-xnli, respectively. These models leveraged the knowledge learned by xlm-roberta, a multilingual variant of Facebook's RoBERTa model released in 2019. It is a large multi-lingual language model, trained on 2.5TB of filtered CommonCrawl data on one hundred languages including Arabic.

The transfer learning process involves a meticulous fine-tuning procedure that was diligently undertaken. The pre-trained models were further trained on Arabic corpus using our Ara-deep dataset for our specific downstream task of DeepFake text detection. In the fine-tuning process, we maintained the architecture and weights of the pre-trained model, incorporated a task-specific classification head for binary classification, and modified the model's parameters utilizing our Ara-deep dataset to tailor the pre-trained knowledge to our specific task (Refer to Section 4.2.2 for the detailed fine-tuning configuration and implementation setup).

This methodological orchestration allowed for a thorough and insightful evaluation of the models' detection capabilities. In Section 4, we provide a detailed presentation of the experiments designed to assess the defensive performance of these detectors.

As shown in Figure 6, the process can be divided into three main phases. The first phase is dedicated to the task of dataset building, the second phase is for fine-tuning pretrained multilingual models, and the third phase is to study the human capacity for distinguishing between DeepFake and genuine text. In general, the second phase is the same for the four employed models. After completing the three phases, the results of both human and automatic detection are analytically compared and evaluated.

Table 4. Hyperparameters Used for Training Our DFTD1, DFTD2, DFTD3, and DFTD4 Models

Hyperparameter	Value
optimizer	AdamW
learning rate	{2e-5, 9e-5, 1e-4, 1e-3}
epochs	{30, 50, 60}
batch size	{10, 16, 32}

The implementation details of the proposed detectors, including the specific fine-tuning hyperparameters (learning rate, number of epochs, batch size) and experimental results, are presented in Section 4.3.

4 Experiments and Results

In this section, we executed a number of experiments in order to evaluate the efficacy of our proposed models. To analyse the performance, an in-depth examination of the results yielded by each individual experiment is conducted. We carried out comparative experiments to evaluate the performance of the proposed models.

4.1 Baseline

Due to the absence of published baseline models for Arabic DeepFake text recognition, this study constructed a **temporal convolution network** (**TCN**) architecture to serve as a baseline for comparison with the proposed models. Refer to Appendix B for the comprehensive technical specifications about the architecture and implementation.

4.2 Experiment I

4.2.1 Objective of the Experiment This experiment seeks to evaluate the efficacy of the constructed models in detecting Arabic DeepFake text utilizing the developed dataset, specifically addressing the initial RQ1 posed in the introduction (RQ1: How effective are the designed models in identifying Arabic DeepFake text using the developed gold standard dataset?).

4.2.2 Setting the Experiment As outlined earlier in Section 3.3, we set up our framework with four multilingual pre-trained models: (1) DFTD1 (mBERT), (2) DFTD2 (XLM-ROBERTA-base), (3) DFTD3 (XLM-ROBERTA-large), and (4) DFTD4 (XLM-ROBERTA-large-xlni). For all experiments, these models underwent optimization with AdamW (Adaptive Moment Estimation) [36], employing a learning rate of 2e-5. We conducted trials over 50 epochs since extending the training beyond this point did not significantly impact the training error. To expedite convergence and reduce memory consumption, a batch size of 10 was adopted during the training process.

Table 4 describes the combinations of hyperparameters used to train the four architectures. The proposed models and their variants are trained and tested utilizing the Ara-Deep Arabic dataset.

To avoid overfitting, we implemented early stopping based on validation performance. To demonstrate the model's capacity for generalization, we report findings on a hold-out validation set. To ensure reproducibility, we randomly divided the dataset into training and testing sets, allocating 80% for training and reserving the remaining 20% for testing. This partitioning remained consistent across all experiments.

During fine-tuning, we meticulously selected optimal hyperparameters through an extensive search encompassing combination of learning rates, batch sizes, and epoch numbers from the following sets {2e-5, 9e-5, 1e-4, 1e-3}, {10, 16, 32}, and {30, 50, 60}, respectively. All of the experiments

Table 5. Results and Performance of the Fine-Tunned Models vs Baseline

Model	Accuracy	F1-score	Precision	recall
DFTD_1	0.9970	0.9970	0.9950	0.9990
DFTD_2	0.9980	0.9980	0.9980	0.9980
DFTD_3	0.9880	0.9881	0.9775	0.9990
DFTD_4	0.9975	0.9975	0.9970	0.9980
TCN	0.8938	0.9100	0.9000	0.9000

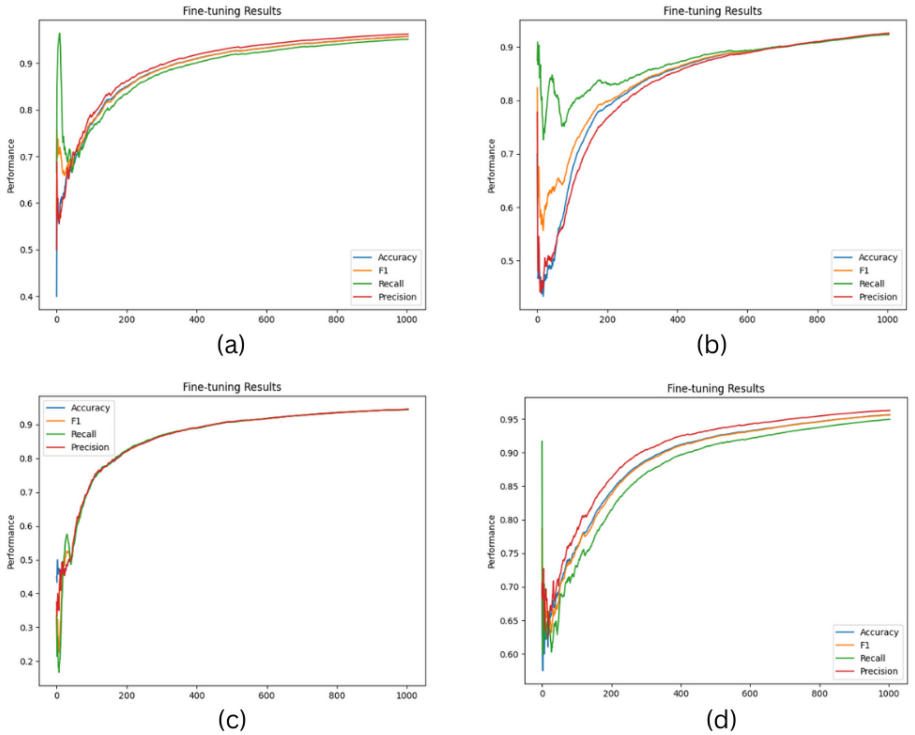

Fig. 7. Performance of the fine-tuned models. Figures (a), (b), (c), and (d) represent the results of evaluating DFTD1, DFTD2, DFTD3, and DFTD4, respectively.

are carried out using Python 3.10.12, transformers 4.33.1, with Cuda 11.8, and trained on a single NVIDIA Tesla T4 GPU.

4.2.3 Results. In this subsection, the results from the experimental investigation I are presented. In addition to the accuracy, the following metrics were also evaluated: Precision, Recall, and F1 score. The precision is the fraction of predicted human-written texts that are actually human-written. The recall is the fraction of actual human-written texts that are predicted to be human-written. The F1 score is the harmonic mean of precision and recall. Table 5 reports the main results of our approach and the comparison with the baseline TCN network's results on the detection task. The accuracy, precision, recall, and F1 scores for each model are shown in Figure 7.

Figure 6 shows the main experimental results of our proposed models, and Table 6 describes in detail the results of the TCN network.

Table 6. Performance Reports of Baseline TCN on Each Class

Label	Precision	Recall	F1-Score	Accuracy
0: Real text	0.9900	0.7200	0.8400	0.8900
1: DeepFake text	0.8000	1.0000	0.8900	

4.2.4 Discussion. This subsection discusses the outcomes of the initial experimental inquiry. The comprehensive experimental results presented in Table 5 indicate that fine-tuning strategies routinely outperform the model trained from scratch. This is attributable to the significantly higher size and increased number of parameters in LLMs.

The differences in the performances of the LLMs and the baseline may stem from the varying capacities of each model to extract structural information from the text:

- —DFTD1: Fine-tuned M(ultilingual)-BERT (179M params): This model displayed outstanding accuracy of 99.70% on the test set. This is a further improvement over the DFTD3 model and suggests that the mBERT base model is able to learn to identify the principal linguistic features that are characteristic of DeepFake text.
- —DFTD2: Fine-tuned XLM-ROBERTA-base (279M params): This model achieved an accuracy of 99.80% on the test set. This is the highest accuracy achieved by any of the models. As can be seen from Table 5, DFTD2 model achieved also the highest precision, and F1 scores. The DFTD2 model is able to learn to identify the most subtle linguistic features of human-written and DeepFake text, and is therefore able to make the most accurate predictions.
- —DFTD3: Fine-tuned XLM-ROBERTA-large (561M params): This model achieved an accuracy of 98.80% on the test set. This is a significant improvement over the baseline accuracy of 91.0%, illustrating the models adeptness in the detection task. The model has demonstrated its aptitude for navigating the challenges presented by managing complex linguistics of AI-generated text.
- —DFTD4: Fine-tuned XLM-ROBERTA-large-xlni (561M params): The performance of DFTD4, which boasted a 99.75% accuracy rate on the test set, was consistent with its enhanced architecture. This model showcased heightened performance, effectively capturing the patterns of ChatGPT-generated text.

The performance analysis of multilingual language models revealed an impressive sensitivity in identifying text produced by ChatGPT. With accuracies rates over 0.98, their cross-lingual prowess underscored the potential as robust detectors in the Arabic language. The culmination of these findings delineates the efficacy of the four models in detecting text originating from ChatGPT. The precision of their classifications highlights their potential for contextually nuanced detection tasks. Analysing the performance of each model, the findings show the nuanced interplay between model architecture, depth, and its performance.

The enlarged architecture of DFTD3 is reflected in its higher sensitivity, supporting the claim that increased model complexity is associated with a stronger ability to grasp contextual complexities. However, DFTD2 outperformed DFTD3 although its architecture is smaller. DFTD2 revealed its high aptitude for accurate text detection by being the best performing model with an accuracy of 99.80%.

DFTD1 remarkable accuracy highlights its innate ability to recognize a wide range of linguistic variations. The fact that DFTD1 (based on the 179M params mBERT) performed better than the much larger DFTD3 was an intriguing outcome. DFTD1 was outperformed by DFTD4 in terms of overall performance.

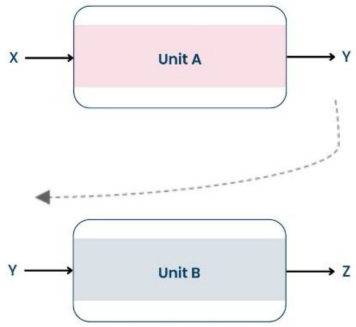

Fig. 8. Rational underpinning experiment II.

Experimental findings demonstrate that all our developed models outperform the baseline method.

4.3 Experiment II

4.3.1 Objective of the Experiment. The aim of this assessment phase is twofold: first, to assess the models' performance, and second, to examine the influence of differing input parameters on decision-making, performance, and the capacity to generalize effectively to unseen examples from diverse domains, sources, and LLMs. This experiment was conducted to investigate the second RQ outlined in the introduction (RQ2: How well do the proposed detection models generalize to unseen data from other benchmark datasets?).

4.3.2 Setting the Experiment. As Figure 8 illustrates, this experiment is designed to examine whether the modification of the input X, the internal architecture of Unit A (the auto-text generation unit), or the output Y generated by Unit A (which serves as the input for Unit B, the detection unit) have a significant influence on the decision processes and overall performance metrics of Unit B (our proposed models). The input X represents the initial data (prompt + text) fed into the LLM, while Unit A is responsible for generating DeepFake text as its output (Y) based on its internal parameters. Unit B, conversely, which encompasses our four proposed detectors: DFTD1, DFTD2, DFTD3, and DFTD4, is responsible for analyzing the outputs from Unit A to determine if they are human-produced or AI-generated.

To validate the performance of our proposed models comprehensively, further evaluation on three different datasets were conducted. These datasets permitted a systematic manipulation of the key variables (X, Unit A, and Y), enabling a comprehensive analysis of their individual and combined effects on the models' behavior. Specifically, we used:

- —M4 dataset: introduced by Wang et al. [55], which serves as a substantial benchmark for machine-generated text detection. Notably, the dataset incorporates a diverse array of examples, ensuring a robust evaluation. In our analysis, we utilized Arabic text generated by ChatGPT. The dataset used for evaluation contains both human text and AI-generated text.
- —LLM Question-Answer Dataset[3]: includes AI-generated texts and prompts produced in 32 different languages by LLMs. The model is given prompts to generate text. The length and complexity of the writings that the LLM produced in response to these prompts are varied. The dataset used for evaluation consists of Arabic DeepFake text generated using the models GPT-3.5, and GPT-4.

[3]https://www.kaggle.com/datasets/trainingdatapro/llm-dataset

Table 7. Evaluation Results on the M4 Dataset

	Accuracy	Precision	Recall	F1-Score
DFTD_1	0.9965	0.9960	0.9980	0.9965
DFTD_2	0.9940	0.9679	0.9970	0.9940
DFTD_3	0.9810	0.9819	0.9800	0.9810
DFTD_4	0.9995	1.0000	0.9990	0.9995

Table 8. Evaluation Results on LLM Question-Answer Dataset

	Accuracy	Precision	Recall	F1-Score
DFTD_1	0.9980	1.0000	0.9960	0.9979
DFTD_2	0.9935	0.9900	0.9970	0.9935
DFTD_3	0.9410	0.8952	0.9990	0.9442
DFTD_4	0.9900	0.9832	0.99700	0.9900

Table 9. Evaluation Results on BLOOM Dataset

Model	Accuracy	Precision	Recall	F1-Score
DFTD_1	0.9935	0.9900	0.9970	0.9935
DFTD_2	0.9940	0.9900	0.9980	0.9940
DFTD_3	0.9950	0.9910	0.9990	0.9950
DFTD_4	0.9555	0.9198	0.9980	0.9573

—BLOOM dataset: We created another dataset for evaluating our models using the BLOOM LLM. The dataset contains both human text and DeepFake text.

4.3.3 Results. The results of our assessment, displayed in Tables 7, 8, and 9, validate the encouraging findings of our proposed models. In the M4 and LLM Question-Answer Datasets, DFTD1 surpassed other models in accuracy, precision, and F1-score, attaining scores of 99.65%, 99.60%, 99.65%, and 99.80%, 100%, and 99.79% respectively. DFTD3 surpasses the other models in both datasets for F1-score, achieving 99.90%.

4.3.4 Discussion. The results underscore their efficacy in not only analysing machine-generated text from diverse sources but also in demonstrating a capacity to generalize proficiently across different domains and LLMs.

According to the logic illustrated in Figure. 8, the assessment of M4 facilitated the validation of the model's performance when Unit A remains constant (Chat-GPT), while X and Y are altered in terms of source and domain. The assessment of the LLM Question-Answer and BLOOM Datasets facilitated the validation of the models' performance when all three critical factors: Unit A (GPT-3.5, GPT-4, and BLOOM), X, and Y deviate from those encountered during training. This experiment demonstrates that the suggested models can effectively generalize to fresh data from diverse areas and sources supplied by various LLMs. This performance across the three evaluation datasets indicates the models' adaptability and highlights their utility.

4.4 Experiment III

4.4.1 Objective of the Experiment. This experiment addresses the third researched RQ (RQ3: How does the proposed detection system compare to the state-of-the-art models in terms of key metrics such as precision, recall, F1-score, and accuracy?), which was raised in the introduction

Table 10. Comparison of Our Best Performing Model on Our Test Data with other State-of-the-Art Detectors of ChatGPT Generations

Model	Accuracy	Precision	Recall	F1_Score
RoBERTa	0.9970	0.9940	1.0000	0.9970
LR-GLTR	0.9740	0.9760	0.9720	0.9740
Stylistic	0.9740	0.9760	0.9720	0.9740
NELA	0.9560	0.9670	0.9430	0.9550
DFTD_4	**0.9990**	**1.0000**	**0.9990**	**0.9990**

section. The aim is to investigate the performance of our proposed model in comparison to other baselines and state-of-the-art models.

4.4.2 Setting the Experiment. Our best performing model in term of accuracy, precision, and F1-score, DFTD4 model, was used in this experiment.

To the best of our knowledge, there are no detectors specifically trained for Arabic DeepFake detection. However, a recent study by Wang et al. [55] focused on detecting ChatGPT-generated text using a diverse dataset that includes multiple languages, including Arabic. Although the study did not explicitly specify detection performance for each language, it provides valuable insights. In our experiment, we compare our top-performing model with the detectors used in the previously mentioned study. This comparison is based on the same dataset M4, which contains Wikipedia of ChatGPT vs Human data, and closely aligns with our custom dataset used for training and testing our models.

4.4.3 Results. In this subsection, the results from the experimental investigation III are presented. The comparison results between our top-performing model (DFTD4) with the detectors from Ref. [55] in term of accuracy, precision, recall, and F1-score are shown in the following table:

4.4.4 Discussion. Overall, our XLM-ROBERTA-large-xnli based model, DFTD4, outperforms the four ChatGPT detectors, the evaluation quantifying the performance of our model relative to state-of-the-art models is encapsulated in Table 10.

Among the existing state-of-the-art models, the highest performance was achieved by the model based on the RoBERTa architecture, yielding an accuracy of 99.70%, as reported in Ref. [55].

A scrutiny of the results presented in Table 10 reveals that our proposed methodology exhibits superior performance and efficacy, as evinced by the higher accuracy of 99.95%, coupled with impressive precision and F1-score of 100% and 99.95% respectively, when juxtaposed with the existing state-of-the-art approaches.

4.5 Experiment IV

4.5.1 Objective of the Experiment. As online content undergoes a profound transformation driven by the widespread proliferation and the unprecedented accessibility of DeepFake text, internet and social media users are increasingly exposed to synthetically generated textual information. In light of this, it becomes imperative to undertake a critical assessment of human capabilities in discerning Arabic DeepFake text. This experiment was designed to tackle this problem by comparing human abilities for detection against auto-detection with the primary objective of this evaluative examination being to assess the judgment abilities of typical internet users. The RQ addressed here is RQ4: How accurately can internet users distinguish between authentic and DeepFake textual content, and how reliable are their judgments?

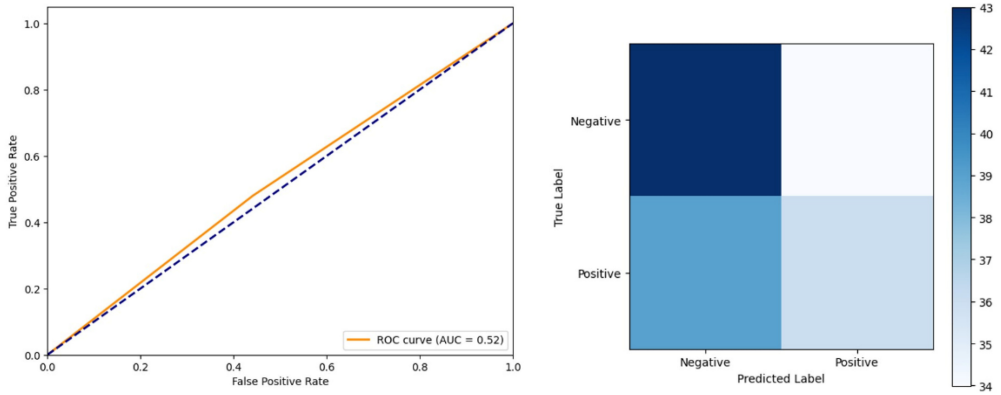

Fig. 9. Visualisation of human performance in the annotation process.

Table 11. Evaluation Results of Several Deepfake Text Detectors in Terms of Accuracy, Precision, Recall, and F1-Score

Detector	Accuracy	precision	Recall	F1-score
DFTD_1	0.9970	0.9970	0.9950	0.9990
DFTD_2	0.9980	0.9980	0.9980	0.9980
DFTD_3	0.9880	0.9881	0.9775	0.9990
DFTD_4	0.9975	0.9975	0.9970	0.9980
Human-detectors	0.5200	0.5100	0.4800	0.5000

4.5.2 Setting the Experiment. To conduct this examination in this experiment, a comprehensive annotation procedure was meticulously crafted, involving the participation of a cohort comprising eight individuals whose native language is Arabic. These human annotators were deliberately devoid of any specialized training in the domain of DeepFake text detection. The annotators were assigned distinct subsets, each comprising 20 data instances drawn from the Ara-Deep dataset, their task encompassed the identification of both DeepFake and authentic text examples within their allocated samples.

4.5.3 Results. Upon collating the feedback provided by all these human annotators, the computed detection accuracy exhibited a mere 51.00%, which was closely resembling random classification with little discriminatory capacity evident in their assessment of text veracity. Figure 9 shows a visualization of human performance in the detection.

ROC Curve: As the classification decision threshold changes, the ROC curve shows the tradeoff between the true positive rate (sensitivity) and the false positive rate. It aids in the visualization of human annotators' performance at various operating points.

AUC: The AUC measures how well human annotators perform overall. It stands for the ROC curve's area under the curve. Better discriminative power is indicated by a larger AUC, with a perfect classifier having an AUC of 1.

The results of human annotation on the Ara-Deep dataset were compared with the gold standard dataset that consists of the true-labeled instances, reflecting the ground truth classifications. Results are illustrated in Table 11 and Figure 10. Table 11 Compares the results of different models used with human performance in terms of the average accuracy, precision, recall, and F1-score values. The best results are highlighted in bold.

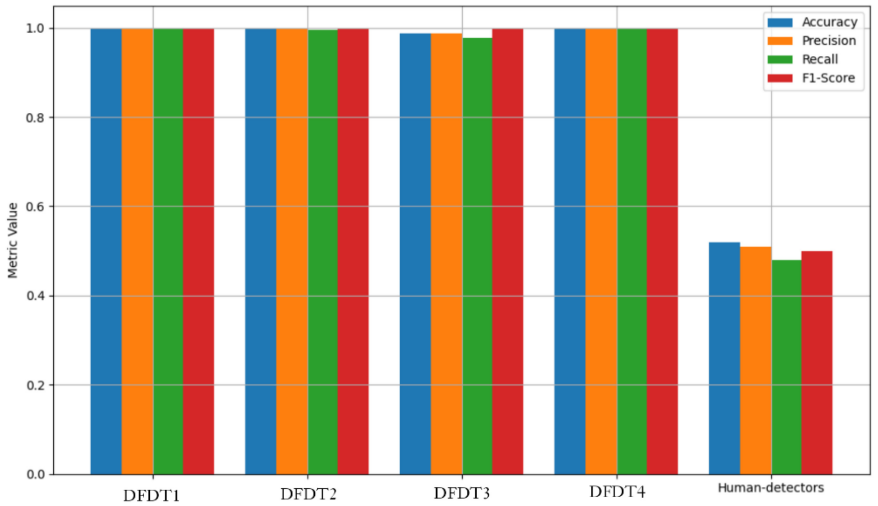

Fig. 10. Performance comparisons of LLM-based DeepFake text detection models with human detection results.

4.5.4 Discussion. The huge difference between the efficacies of the aforementioned models with human detection abilities unveils critical concerns within the realm of LLMs' text detection. The comprehensive analysis of the four models—mBERT, XLM-ROBERTA-BASE, XLM-ROBERTA-LARGE, and XLM-ROBERTA-LARGE-XLNI—evinces their remarkable prowess in identifying text generated by such models. The models' capability to harness intricate contextual relationships and cross-lingual comprehension stands out as a testament to the advancements in DL technology. This suggests that the models are able to learn how to identify the linguistic features of human-written text more effectively than humans.

Nevertheless, the comparison with human detection elicits multifaceted considerations. It is important to recognize that LLM text easily tricks the human cognitive capacity for contextual understanding and nuanced interpretation, even with the human's capability to make subtle inferences beyond the immediate linguistic constructs. However, the models' performance, as evidenced by their high accuracy rates, mirrors their superiority and proves the importance of automating the DeepFake text detection tasks.

Furthermore, while models excel in processing voluminous data and executing repetitive tasks with speed and precision, human intuition and domain knowledge fall short in deciphering complex linguistic nuances and ambiguities between real and coherent LLM generated text when it comes to large amounts of data, especially in lengthy text samples with no syntax or grammar errors.

The LLMs' superior performance over humans may have been influenced by a variety of reasons. The models can process text far more quickly than people can. They can therefore quickly analyse big amounts of data. The second advantage is that the models can also gain knowledge from a large body of data. This indicates that they are able to spot patterns that people would miss. Finally, the models can be fine-tuned on certain datasets, which enables them to be improved on text generated by a particular language model or text relevant to a given area.

The comparison's findings imply that the proposed models are a promising technique for identifying text produced by LLMs. It is crucial to note that the models are not flawless. Nonetheless, they are susceptible to errors, especially when the text is well-written or has a lot of noise. Besides, the quality of the models depends on the data which they were trained on. The models cannot output reliable results if the data is not indicative of reality.

Table 12. Analysis of Typical Sample Cases and Their Classification Results

	Text	Method	Prediction	Label
Example 01	عملية بتوقع حيث استحواذ، صفقات إلى له مقراً سويسرا من ويخلد أبوظبي تملكه الذي فالكون بنك يسعى الذي فيستر وأضاف ،البنك في الخاصة الصيرفة رئيس فيستر، أريك قاله ما وفق المصرفية، الصناعة في اندماجات الأساسية نشاطه أسواق تغطي استحواذ صفقات لعقد مستعداً يكون سوف البنك أن الماضي، يدير في منصبه تولى فيستر وأوضح ،الإماراتية "البيان" صحيفة بحسب آسيا، شرق وجنوب وإفريقيا الشرقية وأوروبا الخليج في المستهدفة من سويسري فرانك ومليارى فرانك مليون 200 بين تتراوح التي البنوك إدارة تشمل التي بالصفقات يتم البنك أن أنبا يخفي ما عليه، تركز وما أهدافها الصفقات تخدم أن بد لا" :ذي في مقابلة في فيستر وقال ،الأصول قيمة حيث ".الرئيسة أسواقها مع متماشية مستهدفة صفقة أي تكون أن بد ولا استراتيجيتنا، في تساهل لن	**DFTD_1**	Real-text	Real-text
		DFTD_2	Real-text	
		DFTD_3	Real-text	
		DFTD_4	Real-text	
		Human-annotators	DeepFake-text	
Example 02	حول العلمية الحقائق تفهم أن يجب العطس، عن "الغريبة" الشعبية المعتقدات بعض فيه تكثر الذي الوقت في مثل أجسامنا، تدخل التي المهيجات من للتخلص لجسمنا إرادية غير استجابة هو العطس ،الطبيعية الظاهرة هذه تنتشر الشعبية المعتقدات بعض أن إلا وشائعًا، طبيعيًا يعتبر العطس أن من الرغم على.القاحية الحبوب أو الغبار أن يعتقد من هناك المثال، سبيل على ،خرافات مجرد وتعتبر علمي أساس لها ليس المعتقدات هذه ومعظم ،حوله لا الافتراضات هذه ،الجسم في الشريرة الأرواح أحد وجود بعني أنه أو الوجه تشوه إلى يؤدي أن يمكن العطس في عادة يتسبب ولا خطيرًا ليس العطس أن هي العلمية الحقيقة.علمي دليل أي إلى تستند ولا بصلة للواقع تمت الجراثيم انتشار لمنع العطس أثناء الأساسية الاحتياطات بعض تتخذ أن ذلك، يجب ومع ،جوهرية صحية مشاكل لمنع وذلك العطس، عند الكوع مرفق أو ورقي بمنديل والأنف الفم يغطي أن يصح ،الشخصية النظافة على والحفاظ لاحظت أو بالعطس مرتبطة مخاوف لديك كانت إذا،المعدية الأمراض انتقال وتقليل المتطايرة الجسيمات انتشار تحجيم يمكنه ،الصحية الرعاية مقدم مع التشاور الأفضل فمن بك، الخاصة العطس عادات في طبيعية غير تغيرات .الصحي وتاريخك أعراضك على بناءً المناسبة المشورة وتوفير حالتك	**DFTD_1**	DeepFake-text	DeepFake-text
		DFTD_2	Real-text	
		DFTD_3	DeepFake-text	
		DFTD_4	DeepFake-text	
		Human-annotators	Real-text	

Overall, this comparison's findings are encouraging. They assert that the models can be used to accurately identify text produced by ChatGPT. However, to increase the models' precision and provide more reliable detection systems, additional study is necessary. Table 12 shows the detection result of each model with the average of human detection on two distinct samples from the two classes (Real and DeepFake text). These instances were extracted from the test set.

5 Practical and Social Implications

The current study aims to examine DeepFake text detection domain in Arabic using LLMs. The study helps to improve our understanding of the linguistic features that distinguish human-written text from machine-generated text. This could lead to the development of more accurate detection systems. Our findings, presented previously, could be examined as fresh knowledge and put to the test using empirical techniques based on different approaches. Additionally, our findings can be expanded in subsequent studies to increase their accuracy and uncover new information about DeepFake text detection.

Additionally, in light of the findings, our results can assist other researchers in defining more specific research objectives to address opportunities and concerns related to the human ability to spot DeepFake text in a more specific and in-depth manner. The findings may be used to create tools supporting the detection of fake news, other forms of misinformation. This may help to protect individuals from being duped or misled by false information. The study's findings could also be used to develop tools for identifying spams and other forms of unwanted content and help to improve the quality of online communication.

From a practical standpoint, the combination of automated and human text detection could function as a robust detector where the collaborative integration of expert-human judgment and algorithmic precision can expedite tasks while retaining nuanced contextual comprehension. Furthermore, the implications extend to real-time text analysis in areas such as online discourse monitoring, ensuring a judicious balance between timely intervention and the contextual intricacies inherent to linguistic communication. These practical insights steer the development of hybrid

systems that capitalize on the strengths of both human expertise and automated algorithms, culminating in an enhanced and efficient framework for language understanding and interpretation.

In the broader societal context, the study aim at addressing issues like misinformation, and disinformation, and to improve the quality of public discourse and decision-making.

The study stresses the analytical study of human potential to mitigate AI-synthesized text and raises concerns about the spread of textual DeepFake content. Other research can build on the study's outcomes to resonate with the societal narrative surrounding AI ethics, engendering discussions on responsible AI integration, equitable decision-making, and the recalibration of societal roles and responsibilities.

6 Conclusions

In this article, an adaptive fine-tuning strategy was applied on a low-resourced language, aimed at detecting Arabic DeepFake text generated by ChatGPT. Our approach employs four distinct models in the detection. The detection system consists of two main parts. Initially, we created a new dataset: Ara-Deep containing authentic and AI-generated text. Then, we performed fine-tuning on mBERT, XLM-ROBERTA-base, XLM-ROBERTA-large, and XLM-ROBERTA-large-xlni. In this work, we conducted multiple experiments with the goal of optimizing hyperparameters and thoroughly exploring the effects of each language model. The experimental results demonstrated the efficacy of the proposed models, which outperformed the baseline DL approach and the state-of-the-art detectors. The study not only advances the understanding of text detection but also contributes to the wider discourse on LLMs performance and adaptability. The comprehensive evaluation of these models serves as a stepping stone toward more robust and nuanced language understanding applications.

In addition, we undertook an evaluation of human capability in distinguishing real text from AI-synthesized text and compared this to auto-detection. The results of this study show that while regular internet users with no prior training often struggle to detect text generated by ChatGPT, LLMs prove highly effective in this task. The detection is extremely important because it can be used to prevent the spread of misinformation and disinformation, which are often generated by LLMs. Finally, we anticipate that this study facilitates the comprehensive and discerning evaluation of the detection capabilities, contributing to the academic discourse on language-model-generated content detection.

7 Limitations, Challenges, and Future Work

The detection of text generated by LLMs represents a critical area of research due to its potential implications for trust, authenticity, and the responsible use of AI-generated content. In this article, we introduce a robust detection system based on LLMs. However, it is important to acknowledge certain limitations and uncovered aspects. Our generated dataset is considered relatively small, primarily due to constraints related to computing resources. Additionally, our DeepFake corpora is generated exclusively by ChatGPT, and does not encompass other state-of-the-art generative models. Furthermore, we did not conduct additional partitioning or randomness tests on the dataset, which are essential for ensuring result reproducibility and mitigating the influence of random factors.

Future research directions involve the examination of the performance of our proposed models across various languages using a larger dataset. The identification of Arabic DeepFake text in domain-specific corpus also, such as education, for example, is of great interest. Significantly, there is substantial room for advancement in countering synthetic adversarial samples that pose challenges to the robustness of detection systems. Addressing this inherent adversarial nature calls for the exploration of game-theoretic frameworks, where the detection model and the text

generation model engage in a dynamic interplay. This, in turn, necessitates the development of iterative training algorithms to ensure ongoing model improvement. Additionally, more effort needs to be made for creating tools that may be applied to enhance communication and information quality and safeguard users from being misled by inaccurate information.

In this study, we focused on the detection of AI-generated text without considering contextual factors. Future research should expand upon these findings by covering context-aware detection methodologies. Additionally, a comprehensive comparative analysis between the detection of in-context and out-of-context text generations would significantly contribute to the field's understanding of AI-generated content identification.

The study's overall findings are encouraging and indicate that LLMs may be utilized to successfully recognize text produced by ChatGPT. However, more research efforts are needed to confirm these findings and to develop more effective detection systems.

References

[1] H. Alamleh, A. A. S. Alqahtani, and A. Elsaid. 2023. Distinguishing human-written and ChatGPT-generated text using machine learning. *2023 Systems and Information Engineering Design Symposium*, SIEDS 2023, 154–158. DOI: 10.1109/SIEDS58326.2023.10137767

[2] Z. Alyafeai, M. S. AlShaibani, and I. Ahmad. 2020. A survey on transfer learning in natural language processing. arXiv preprint arXiv:2007.04239. Retrieved from https://arxiv.org/abs/2007.04239

[3] W. Antoun, F. Baly, and H. Hajj. 2020. AraBERT: Transformer-based model for arabic language understanding. arXiv preprint arXiv:2003.00104. Retrieved from https://arxiv.org/abs/2003.00104

[4] W. Antoun, V. Mouilleron, B. Sagot, and D. Seddah. 2023. Towards a robust detection of language model generated text: Is ChatGPT that easy to detect? arXiv preprint arXiv:2306.05871. Retrieved from https://arxiv.org/abs/2306.05871

[5] P. Azunre. 2021. Transfer learning for natural language processing. *Simon and Schuster*. New York, NY.

[6] G. Bharathi Mohan, R. Prasanna Kumar, N. L. Keerthana, D. Mukesh, R. M. Hemesh, I. V. Priyanka, and S. Parthasarathy. 2023. Cross-lingual machine translation: An analysis model for low resource languages. In *International Conference on Emerging Trends and Technologies on Intelligent Systems. Singapore: Springer Nature Singapore.* 81–94. DOI: 10.1007/978-981-99-3963-3_7

[7] A. Bhattacharjee and H. Liu. 2024. Fighting fire with fire: Can ChatGPT detect AI-generated text?. *ACM SIGKDD Explorations Newsletter* 25, 2 (2024), 14–21.

[8] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, and D. Amodei. 2020. Language models are few-shot learners—special version. *Conference on Neural Information Processing Systems (NeurIPS 2020)*, (NeurIPS), 1–25.

[9] G. Bueermann and N. Perucica. 2023. How can we combat the worrying rise in deepfake content?', *World Economic Forum*, 1 January, Retrieved from https://www.weforum.org/agenda/2023/05/how-can-we-combat-the-worrying-rise-in-deepfake-content/

[10] R. Cao, Y. Wang, L. Gao, and M. Yang. 2023. DictPrompt: Comprehensive dictionary-integrated prompt tuning for pre-trained language model. *Knowledge-Based Systems*, 273, 110605. DOI: 10.1016/j.knosys.2023.110605

[11] J. E. Casal and M. Kessler. 2023. Can linguists distinguish between ChatGPT/AI and human writing?: A study of research ethics and academic publishing. *Research Methods in Applied Linguistics* 2, 3 (2023), 100068. DOI: 10.1016/j.rmal.2023.100068

[12] C. Chaka. 2023. Detecting AI content in responses generated by ChatGPT, YouChat, and Chatsonic: The case of five AI content detection tools. *Journal of Applied Learning and Teaching* 6, 2 (2023).

[13] S. Chakraborty, A. S. Bedi, S. Zhu, B. An, D. Manocha, and F. Huang. 2023. On the possibilities of ai-generated text detection. arXiv preprint arXiv:2304.04736. Retrieved from https://arxiv.org/abs/2304.04736

[14] Y. Chen, M. Wen, K. Zhang, and S. Yu. 2020. Short term photovoltaic output prediction based on similar day matching and TCN attention. *Electr. Meas. Instrum*, 1–9.

[15] R. Collobert and J. Weston. 2008. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th International Conference on Machine Learning*. 160–167.

[16] A. Conneau et al. 2020. Unsupervised cross-lingual representation learning at scale. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. 8440–8451. DOI: 10.18653/v1/2020.acl-main.747

[17] J. D. M. W. C. Kenton and L. K. Toutanova. 2019. 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 (Mlm), 4171–4186.

[18] Y. Du, Q. Li, L. Wang, and Y. He. 2020. 'Biomedical-domain pre-trained language model for extractive summarization'. *Knowledge-Based Systems* 199 (2020), 105964. DOI: 10.1016/j.knosys.2020.105964

[19] O. Einea, A. Elnagar, and R. Al Debsi. 2019. "SANAD: Single-label arabic news articles dataset for automatic text categorization. *Data in Brief* 25 (2019), 104076. DOI: 10.1016/j.dib.2019.104076

[20] A. M. Elkhatat, K. Elsaid, and S. Almeer. 2023. Evaluating the efficacy of AI content detection tools in differentiating between human and AI-generated text. *International Journal for Educational Integrity* 19, 1 (2023), 17.

[21] T. Fagni, F. Falchi, M. Gambini, A. Martella, and M. Tesconi. 2021. TweepFake: About detecting deepfake tweets. *PLoS One* 16 (5 May), 1–19. DOI: 10.1371/journal.pone.0251415

[22] L. Fan, L. Li, Z. Ma, S. Lee, H. Yu, and L. Hemphill. 2024. A bibliometric review of large language models research from 2017 to 2023. *ACM Transactions on Intelligent Systems and Technology* 15, 5 (2024), 1–25.

[23] W. Fu et al. 2021. Output-based transfer learning in genetic programming for document classification. *Knowledge-Based Systems*, 212 (2021), 106597. DOI: 10.1016/j.knosys.2020.106597

[24] J. Guan, F. Huang, Z. Zhao, X. Zhu, and M. Huang. 2020. A knowledge-enhanced pretraining model for commonsense story generation. *Transactions of the Association for Computational Linguistics* 8 (2020), 93–108. DOI: 10.1162/tacl_a_00302

[25] F. Harrag and M. K. Djahli. 2022. Arabic fake news detection: A fact checking based deep learning approach. *ACM Transactions on Asian and Low-Resource Language Information Processing* 21, 4 (2022), 1–34. DOI: 10.1145/3501401

[26] F. Harrag, M. Debbah, K. Darwish, and A. Abdelali. 2021. BERT transformer model for detecting arabic GPT2 auto-generated tweets. arXiv preprint arXiv:2101.09345. Retrieved from https://arxiv.org/abs/2101.09345

[27] M. Hoang, O. Alija Bihorac, and J. Rouces. 2019. Aspect-based sentiment analysis using BERT. *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, 187–196. Available at: https://aclanthology.org/W19-6120/

[28] R. Kumar and M. Mindzak. 2023. Distinguishing human generated text from ChatGPT generated text using machine learning. *Available at*: http://arxiv.org/abs/2306.01761

[29] J. Vincent. 2018. Why we need a better definition of 'deepfake'. *The Verge, 22.. (Accessed on 2/15/2023)*. Retrieved from https://www.theverge.com/2 018/5/22/17380306/deepfake-definition-ai-manipulation-fake -news

[30] I. Katib, F. Y. Assiri, H. A. Abdushkour, D. Hamed, and M. Ragab. 2023. Differentiating chat generative pretrained transformer from humans: Detecting ChatGPT-generated text and human text using machine learning. *Mathematics* 11, 15 (2023), 3400.

[31] K. Krishna, Y. Song, M. Karpinska, J. Wieting, and M. Iyyer. 2024. Paraphrasing evades detectors of AI-generated text. *But Retrieval is an Effective Defense'*, (NeurIPS), 1–32.

[32] T. Kumarage, J. Garland, A. Bhattacharjee, K. Trapeznikov, S. Ruston, and H. Liu. 2023. Stylometric detection of AI-generated text in twitter timelines, 1–13. Available at: http://arxiv.org/abs/2303.03697

[33] C. Lea, M. D. Flynn, R. Vidal, A. Reiter, and G. D. Hager. 2017. Temporal convolutional networks for action segmentation and detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 156–165.

[34] B. Li, Y. He, and W. Xu. 2021. Cross-lingual named entity recognition using parallel corpus: A new approach using XLM-RoBERTa alignment. arXiv preprint arXiv:2101.11112. Retrieved from https://arxiv.org/abs/2101.11112

[35] Y. Liu. 2019. RoBERTa: A robustly optimized BERT pretraining approach, (1). Available at: http://arxiv.org/abs/1907.11692

[36] I. Loshchilov and F. Hutter. 2019. Decoupled weight decay regularization. *7th International Conference on Learning Representations*, ICLR 2019.

[37] C. Mao, J. Xu, L. Rasmussen, Y. Li, P. Adekkanattu, J. Pacheco, B. Bonakdarpour, R. Vassar, L. Shen, G. Jiang, F. Wang, J. Pathak, and Y. Luo. 2023. AD-BERT: Using pre-trained language model to predict the progression from mild cognitive impairment to Alzheimer's disease'. *Journal of Biomedical Informatics*, 144(July), 104442. DOI: 10.1016/j.jbi.2023.104442

[38] A. Martín, J. Huertas-Tato, Á. Huertas-García, G. Villar-Rodríguez, and D. Camacho. 2022. FacTeR-Check: Semi-automated fact-checking through semantic similarity and natural language inference. *Knowledge-Based Systems*, 251, 109265. DOI: 10.1016/j.knosys.2022.109265

[39] M. Nicholson, R. Agrahari, C. Conran, H. Assem, and J. D. Kelleher. 2022. The interaction of normalisation and clustering in sub-domain definition for multi-source transfer learning based time series anomaly detection. *Knowledge-Based Systems*, 257, 109894. DOI: 10.1016/j.knosys.2022.109894

[40] J. Pan, X. Hu, Y. Zhang, P. Li, Y. Lin, H. Li, W. He, and L. Li. 2015. Quadruple transfer learning: Exploiting both shared and non-shared concepts for text classification. *Knowledge-Based Systems* 90 (2015), 199–210. DOI: 10.1016/j.knosys.2015.09.017

[41] S. J. Pan and Q. Yang. 2010. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering* 22, 10 (2010), 1345–1359. DOI: 10.1109/TKDE.2009.191

[42] R. Pappagari, P. Zelasko, J. Villalba, Y. Carmiel, and N. Dehak. 2019. Hierarchical transformers for long document classification. *2019 IEEE Automatic Speech Recognition and Understanding Workshop, ASRU 2019—Proceedings*, 838–844. DOI: 10.1109/ASRU46091.2019.9003958

[43] C. Qu et al. 2019. BERT with history answer embedding for conversational question answering. *SIGIR 2019—Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1133–1136. DOI: 10.1145/3331184.3331341

[44] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and J. L. P. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research* 21, 140 (2020), 1–67.

[45] S. Ruder, M. E. Peters, S. Swayamdipta, and T. Wolf. 2019. Transfer learning in natural language processing tutorial. *NAACL HLT 2019–2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies–Tutorial Abstracts*, (2010), 15–18.

[46] V. S. Sadasivan, A. Kumar, S. Balasubramanian, W. Wang, and S. Feizi. 2023. Can AI-generated text be reliably detected?. arXiv preprint arXiv:2303.11156. Retrieved from https://arxiv.org/abs/2303.11156

[47] V. Sanh. 2019. DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter'. 2–6. arXiv preprint arXiv:1910.01108. Retrieved from https://arxiv.org/abs/1910.01108

[48] S. Shi, E. Zhao, D. Tang, Y. Wang, P. Li, W. Bi, H. Jiang, G. Huang, L. Cui, X. Huang, C. Zhou, Y. Dai, and D. Ma. 2022. Effidit: Your AI writing assistant'. arXiv preprint arXiv:2208.01815. Retrieved from https://arxiv.org/abs/2208.01815

[49] A. Srinivasan, S. Sitaram, T. Ganu, S. Dandapat, K. Bali, and M. Choudhury. 2021. Predicting the performance of multilingual NLP models. arXiv preprint arXiv:2110.08875. Retrieved from https://arxiv.org/abs/2110.08875

[50] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M. A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample. 2023. LLaMA: Open and efficient foundation language models. Available at: http://arxiv.org/abs/2302.13971

[51] L. Uzun. 2023. ChatGPT and academic integrity concerns: Detecting artificial intelligence generated content. *Language Education and Technology* 3, (2023) 1.

[52] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems (NIPS'17)*. 5998–6008. https://doi.org/10.48550/arXiv.1706.03762

[53] V. Verma, E. Fleisig, N. Tomlin, and D. Klein. 2023. Ghostbuster: Detecting text ghostwritten by large language models. 1–18. Retrieved from http://arxiv.org/abs/2305.15047

[54] J. Wang, X. Zhang, and L. Chen. 2021. How well do pre-trained contextual language representations recommend labels for GitHub issues?. *Knowledge-Based Systems*, 232 (2021), 107476. DOI: 10.1016/j.knosys.2021.107476

[55] Y. Wang et al. 2023. M4: Multi-generator, multi-domain, and multi-lingual black-box machine-generated text detection. Available at: http://arxiv.org/abs/2305.14902

[56] Z. Wang et al. 2019. Multi-passage BERT: A globally normalized BERT model for open-domain question answering. *EMNLP-IJCNLP 2019–2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, (1), 5878–5882. DOI: 10.18653/v1/d19-1599

[57] D. Weber–Wulf, A. Anohina–Naumeca, S. Bjelobaba, T. Foltýnek, J. Guerrero–Dib, O. Popoola, P. Šigut, and L. Waddington. 2023. Testing of detection tools for AI-generated text. *International Journal for Educational Integrity* 19, 1 (2023), 26.

[58] K. Wu, L. Pang, H. Shen, X. Cheng, and T. S. Chua. 2023. LLMDet: A large language models detection tool. arXiv preprint arXiv:2305.15004. Retrieved from https://arxiv.org/abs/2305.15004

[59] S. Xie et al. 2021. PALI at SemEval-2021 Task 2: Fine-tune XLM-RoBERTa for word in context disambiguation. *SemEval 2021–15th International Workshop on Semantic Evaluation, Proceedings of the Workshop*, 713–718. DOI: 10.18653/v1/2021.semeval-1.93

[60] G. Xu et al. 2022. Aspect-level sentiment classification based on attention-BiLSTM model and transfer learning. *Knowledge-Based Systems*, 245, (2022), 108586. DOI: 10.1016/j.knosys.2022.108586

[61] Z. Xu, R. Xu, and V. S. Sheng. 2024. ChatGPT-Generated code assignment detection using perplexity of large language models (student abstract). In *Proceedings of the AAAI Conference on Artificial Intelligence* 38, 21 (2024), 23688–23689).

[62] R. Touma, H. Hajj, W. El-Hajj, and K. Shaban. 2023. Automated generation of human-readable natural Arabic text from rdf data. *ACM Transactions on Asian and Low-Resource Language Information Processing* 22, 4 (2023), 1–13.

[63] S. Demir. 2022. Turkish data-to-text generation using sequence-to-sequence neural networks. *ACM Transactions on Asian and Low-Resource Language Information Processing* 22, 2 (2022), 1–27.

[64] A. Kumar and V. H. C. Albuquerque. 2021. Sentiment analysis using XLM-R transformer and zero-shot transfer learning on resource-poor indian language. *Transactions on Asian and Low-Resource Language Information Processing* 20, 5 (2021), 1–13.

[65] M. Bozuyla. 2024. Sentiment analysis of turkish drug reviews with bidirectional encoder representations from transformers. *ACM Transactions on Asian and Low-Resource Language Information Processing* 23, 1 (2024), 1–17.

[66] H. Al-Omari and R. Duwairi. 2023. So2al-wa-Gwab: A new arabic question-answering dataset trained on answer extraction models. *ACM Transactions on Asian and Low-Resource Language Information Processing* 22, 8 (2023), 1–21.

[67] Z. Dong, J. Ni, D. M. Bikel, E. Alfonseca, Y. Wang, C. Qu, and I. Zitouni. 2022. Exploring dual encoder architectures for question answering. arXiv:2204.07120. Retrieved from https://arxiv.org/abs/2204.07120

[68] M. Khalilia, S. Malaysha, R. Suwaileh, M. Jarrar, A. Aljabari, T. Elsayed, and I. Zitouni. 2024. ArabicNLU 2024: The first arabic natural language understanding shared task. arXiv:2407.20663. Retrieved from https://arxiv.org/abs/2407.20663

[69] X. Liu and L. Kong. 2024. AI text detection method based on perplexity features with strided sliding window. *Working Notes of Clef*.

[70] T. Kumarage, G. Agrawal, P. Sheth, R. Moraffah, A. Chadha, J. Garland, and H. Liu. 2024. A survey of ai-generated text forensic systems: Detection, attribution, and characterization. arXiv:2403.01152. Retrieved from https://arxiv.org/abs/2403.01152

[71] Y. Mo, H. Qin, Y. Dong, Z. Zhu, and Z. Li. 2024. Large language model (llm) ai text generation detection based on transformer deep learning algorithm. arXiv:2405.06652. Retrieved from https://arxiv.org/abs/2405.06652

[72] Z. Lai, X. Zhang, and S. Chen. 2024. Adaptive ensembles of fine-tuned transformers for llm-generated text detection. arXiv:2403.13335. Retrieved from https://arxiv.org/abs/2403.13335

Appendix A

Dataset Creation

The dataset used in our research is Ara-Deep, a corpus that we have generated to train our DeepFake text detection models. The initial phase of our dataset generation process involves the acquisition of authentic text data, which serves as the foundation for generating DeepFake text using the popular version of ChatGPT. The authentic text corpus was sourced from the publicly accessible SANAD (Single-Label Arabic News Articles Dataset) corpora on Kaggle [19]; specifically, we used the corpus of Arabic articles that was compiled from thousands of articles extracted from the online newspaper Al-Arabiya. This corpus encompasses seven distinct categories, namely: religion, culture, technology, politics, sports, finance, and medical. We use the authentic text corpus dataset by systematically collecting 160 articles from each category, with the exclusion of specific categories (Political and Religion) to prevent potential bias in the generated text.

Prompts Engineering

Prompt engineering refers to the process of creating and refining the input authored by a human and provided to generate text by a LLM. The prompt is a textual input that informs the LLM on the sequence of words to generate as a text output that effectively communicates the desired end of the user. Since language models have a high potential for natural language production, prompt engineering is considered particularly crucial for LLMs.

In our study, the purpose of prompt engineering is to develop prompts that are effective at getting ChatGPT to generate consistent, fluent, coherent, and human-like natural Arabic writing. The prompts utilized ranged from simple as a single sentence to more complicated prompts that included directions and illustrations. This procedure enabled the created outputs to be narrowed to match the genuine text style, boosting the overall accuracy and relevancy of the produced text.

To that end, we used articles from five different categories in SANAD. From each category, 160 articles were used as input. Given the real text x as input, ChatGPT outputs the completion or reformulation y depending on the current prompt.

The articles employed in the prompt engineering for generating ChatGPT DeepFake text were deliberately omitted from incorporation into the definitive constructed dataset. Instead, the authentic textual corpus within Ara-deep encompasses the same number of other articles from the aforementioned categories, sourced from SANAD. This procedure is undertaken to ensure the establishment of a balanced and comprehensive dataset.

Appendix B

Baseline

Temporal convolution network (TCN): TCNs [33] are a type of convolutional neural network (CNN) that are specifically designed for processing sequential data. They achieve this by using a hierarchy of temporal convolutions, which allow them to learn long-range temporal

Table 13. Hyperparameters for Our TCN Baseline Model.1

Hyperparameter	Value	
Embeddings	Ara-BERT	
Filters_size	Layer_1	(1, 2, 4)
	Layer_2	(1, 2, 4)
Num_filters	Layer_1	128
	Layer_2	64
Epochs	50	
Dense_units	17	

dependencies. In addition to flexible sensing field size, low memory usage, and parallel computing, the unique TCN network topology also could prevent gradient disappearance or explosion during RNN training [14]. The TCN model uses the Ara-BERT language model [3] for generating embeddings and adds additional layers on top for classification. The added layers include: (1) SpatialDropout1D Layer: This layer applies dropout regularization to the input tensor, (2) TCN Layers: There are two TCN (Temporal Convolutional Network) layers employed, with the first layer having 128 filters and (1, 2, 4) dilations and the second layer having 64 filters and (1, 2, 4) dilations, (3) the two pooling layers GlobalAveragePooling1D and GlobalMaxPooling1D are used, (4) Concatenate Layer: This layer concatenates the results of the pooling layers, (5) Dense Layers: One dense layer with 16 units is used, (6) Dropout Layer: This layer applies dropout regularization to the input tensor, and (7) Output Layer: One dense layer with one unit and sigmoid activation is used for binary classification. The hyperparameters used when implementing the TCN network are summarized in Table 13.

Received 12 September 2024; revised 26 November 2024; accepted 8 December 2024