AI Text Detectors and the Misclassification of Slightly Polished Arabic Text

Saleh Almohaimeed[1], Saad Almohaimeed[2], Mousa Jari[1], Khaled A. Alobaid[1], Fahad Alotaibi[1]

[1]College of Applied Computer Science, King Saud University, Riyadh, Saudi Arabia.
[2]Department of Digital Transformation, Institution of Public Administration, Riyadh, Saudi Arabia.

Contributing authors: salmohaimeed@ksu.edu.sa; mohaimeeds@ipa.edu.sa; Maljari@ksu.edu.sa; Kalobaid@ksu.edu.sa; faalotaibi@ksu.edu.sa;

Abstract

Many AI detection models have been developed to counter the presence of articles created by artificial intelligence (AI). However, if a human-authored article is slightly polished by AI, a shift will occur in the borderline decision of these AI detection models, leading them to consider it as AI-generated article. This misclassification may result in falsely accusing authors of AI plagiarism and harm the credibility of AI detectors. In English, some efforts were made to meet this challenge, but not in Arabic. In this paper, we generated two datasets. The first dataset contains 800 Arabic articles, half AI-generated and half human-authored. We used it to evaluate 14 Large Language models (LLMs) and commercial AI detectors to assess their ability in distinguishing between human-authored and AI-generated articles. The best 8 models were chosen to act as detectors for our primary concern, which is whether they would consider slightly polished human-authored text as AI-generated. The second dataset, Ar-APT, contains 400 Arabic human-authored articles polished by 10 LLMs using 4 polishing settings, totaling 16400 samples. We use it to evaluate the 8 nominated models and determine whether slight polishing will affect their performance. The results reveal that all AI detectors incorrectly attribute a significant number of articles to AI. The best performing LLM, Claude-4 Sonnet, achieved 83.51%, its performance decreased to 57.63% for articles slightly polished by LLaMA-3. Whereas the best performing commercial model, originality.AI, achieves 92% accuracy, dropped to 12% for articles slightly polished by Mistral or Gemma-3.

Keywords: Arabic AI Detection, , AI-polished text, Detection fairness

1 Introduction

Recent years have seen a great deal of competition between Large Language Model (LLM) companies to provide better multilingual models with an advanced level of capabilities. One of their abilities is to generate text that cannot be distinguished from authentic human writing. Due to this, a large number of English text-detector models have been developed. These detectors are designed to distinguish human-authored text from AI-generated content. However, a critical issue arises when someone uses LLM to slightly polish their own human writing without changing its meaning or adding additional information, as shown in Figure 1. In such cases, text detectors struggle to distinguish whether it is a human-written or entirely generated by artificial intelligence. Consequently, there is a risk of falsely accusing the article of being generated by artificial intelligence when it is not. According to Gillham (2025), 10.86% of articles on 246 Fortune 500 companies' blogs are generated by artificial intelligence. However, this may not be true, as authors may only use LLM to enhance the readability of their articles slightly.

Fig. 1 This is an example of an Arabic human-written article that has been polished twice under 10% and 50% polishing settings. Words with red colors are polished words in comparison to the human-written text

From Saha and Feizi (2025), the authors study this issue and provide clear explanations as to why better models are needed to solve this problem. However, the study focused only on English, and did not give much consideration to other languages, such as Arabic. Compared to many other languages, the Arabic language has a richer morphology and a variety of dialects. Moreover, the use of similar words with different diacritics may lead to a different meaning, even if they are within the same context. Having to deal with these issues makes it more difficult to build a text detector for Arabic articles.

In this paper, we have conducted a comprehensive investigation using more than 300 experiments in order to address three main concerns. One, what is the effectiveness of popular LLMs and commercial AI models in distinguishing between Arabic human-written articles and those generated by AI. Second, we investigate the effectiveness of these LLMs in terms of polishing text without changing the meaning. Third, we retest the LLMs and commercial AI models to determine whether they consider slightly polished Arabic text as AI-generated.

In order to accomplish this, we first constructed two datasets. The first dataset consists of 800 samples, half of which are human-written and the other half are generated by AI. The second dataset is Ar-APT [1] (Arabic AI polish text), which contains 400 authentic Arabic human texts polished by 10 LLMs, each with four different polishing percentage levels, resulting in 16400 samples.

The first dataset was used as a means to address our first concern, which is testing the ability of LLMs and very popular commercial models to differentiate between human-authored articles and AI-generated articles. The second dataset Ar-APT is used to address the aforementioned second and third concerns. We also used Ar-APT to demonstrate our main objective, which is the problem of falsely accusing human-written articles as AI-generated when only slight polishing was done.

There were 14 experiments conducted on the first dataset using different LLMs, such as the GPT-4o OpenAI (2024), LLaMA-3 70B Meta (2024), and Claude-4 Sonnet Anthropic (2024), as well as commercial models, such as ZeroGPT Baroud and Oraginality-AI Gillham. We will nominate models that achieve acceptable results to act as text detectors on the Ar-APT dataset. After that, we conducted 320 experiments on the Ar-APT dataset to determine whether nominated LLMs and Commercial models consider human-polished text to be AI-generated or not.

As a result of our findings, it is necessary to design models specifically developed to serve as text detectors for Arabic texts. As the best performing publicly available model is Claude-4 Sonnet Anthropic (2024), which achieved an accuracy of 83.51% and a false positive rate of 16.49%, while the best commercial model achieved 96% accuracy and 8% false positive rate.

Moreover, on Arabic slightly polished text, all experimented LLMs and commercial models experienced a drop in performance in determining whether the text was generated using AI or not. Under 10% polishing percentage, the Claude-4 Sonnet Anthropic (2024) produces an accuracy rate of 65.03% and a false positive rate of 34.97% with articles slightly polished by Mistral Mensch (2024). Moreover, with articles that were polished by Gemma-3 DeepMind (2025), Claude-4 Sonnet got 65.71%

[1] github.com/Saleh-Almohaimeed/Ar-APT

accuracy, 34.29% false positive rate. Thus, with only slightly polished articles, Claude-4 Sonnet's accuracy decreased by approximately 18%.

Furthermore, the best performing commercial model, Originality.AI Gillham, exhibited an impressive performance decline when detecting human articles that were slightly polished (10%) by Mistral Mensch (2024) and Gemma-3 DeepMind (2025). Originality.AI accuracy has dropped to 12% and the false positive rate jumped to 88%. This clearly highlights the necessity of addressing this issue, both for researchers and for the owners of the popular Originality.AI Gillham model.

Moreover, an analysis was conducted to determine whether LLMs are capable of polishing arabic text while keeping the meaning unchanged. In the results, Claude-4 Sonnet performed best, followed by Deepseek-3.1 Deepseek (2025) and GPT-4o OpenAI (2024). Additionally, all of the 10 LLMs used for polishing managed to maintain the meaning and got a high cosine similarity score, except for the Qwen-3 Alibaba (2025) model, which appears to have difficulty with Arabic texts. Contributions:

1. Constructed the first dataset that contains 800 samples of AI-generated text and human-authored text.
2. A second dataset (Ar-APT) was constructed using 400 human-authored texts and an additional 16000 AI-polished versions of these texts, generated by 10 LLMs, under 4 degrees of polishing.
3. 14 experiments were conducted to address our first concern about how well popular LLMs and commercial models distinguish between Arabic human-written articles and AI-generated ones.
4. Our second concern was addressed by calculating the cosine similarity scores and the Jaccard similarity scores between all polished articles and their original counterparts. In order to determine how effective LLMs are at polishing text without altering the meaning.
5. 320 experiments were undertaken to answer our third concern, to determine whether LLMs and commercial models consider slightly polished Arabic text to be AI-generated.

2 Related Works

As technology has become increasingly present in all aspects of people's lives. Information is now shared on social media platforms, blog websites, product reviews, and many other places. Consequently, a number of issues have emerged from the misuse of technology. As artificial intelligence tools, particularly deep learning tools, are becoming more prevalent, it has become increasingly important to address these issues and find appropriate solutions. For example, to combat hateful content spread across social media, an NLP task called Hate Speech detection Almohaimeed (2023),Almohaimeed (2024a) was designed to identify hateful content. Similarly, text bias detection models Farrelly and Singh (2023),Raza et al. (2023) have been introduced for articles that show bias toward certain races or cultures. An additional instance of technology being misused involves using AI models to generate articles and then falsely claiming that they were written by humans. A new method of addressing this issue has emerged

called AI-generated content detection Chaka (2024),Alshammari et al. (2024), which has led to the development of numerous free and paid models.

In order to finetune these models and distinguish between human-written text and AI-generated text, we need a proper dataset. A number of datasets were released, including M4 Wang et al. (2024b) and RAID Dugan et al. (2024). RAID is not a multilingual dataset, it is only in English. It includes 6 million generated content across 11 models and 8 domains. M4 Wang et al. (2024b) is a noteworthy dataset that spans multiple domains and includes AI-content generated by 8 LLMs. A total of 7 languages were represented in the dataset, including English, Chinese, Urdo, Indonesian, Bulgarian, and Arabic. Additionally, there are several datasets which only support English, such as the MAGE Li et al. (2024) dataset, and the HC-Var Xu et al. (2024) dataset. There are also a number of datasets that support a wide variety of languages, such as HC3-Plus Su (2024), and AuText2023 Sarvazyan et al. (2023). In our paper, we built our own dataset for two reasons: in order to ensure diversity, we wanted the Arabic AI Text to be generated by more LLMs, so 10 LLMs were utilized. Secondly, we want to ensure that the human-authored text is sourced from a variety of sources. The majority of datasets focus on news-only resources; however, ours is collected from popular news and blog sources.

Some of the traditional methods used to detect AI-generated articles such as perplexity Gehrmann et al. (2019), which emphasize that the LLM generally chooses the highly frequent next word when writing the AI article. Hence, if the next words are mostly too obvious, they flag them as likely to be generated by AI. In a more advanced approach than perplexity Gehrmann et al. (2019), models such as BERT Wang et al. (2024a) and XLM-Roberta Macko et al. (2023) have been fine-tuned to act as AI detectors. The main purpose of these models is to differentiate between pure artificial intelligence and human-written text. However, if you are including human-written text that has been slightly polished by AI in a way that does not change the meaning, there are no models designed for this purpose. Some studies have addressed the problem without developing AI detecting models, such as Saha and Feizi (2025). The authors constructed a dataset of 300 samples, then polished the 300 samples using 5 LLMs with 11 polishing settings, producing approximately 15K samples. Each original human-written sample is polished 11 times. They then evaluated 12 detectors and demonstrated that even slight polishing could affect the detector decision, falsely reporting the article was generated entirely by AI. In our paper, we did the same; however, instead of 5 LLMs, we used 10 LLMs for polishing the Arabic article. After we realized that the 11 polishing settings used in their study Saha and Feizi (2025) [1%, 5%, 10%, 20%, 35%, 50%, 75%, extremely minor, minor, slightly major, and major] did not significantly impact detector performance, we decided to stick to four polishing settings [10%, 25%, 50%, 75%]. In this way, we generated the same number of samples using more LLMs and fewer polishing settings.

In contrast, very few efforts have been made for the Arabic language, such as Al-Shaibani and Ahmed (2025), which has tested its own dataset of texts extracted from the abstract Arabic research papers and social media reviews. They used 4 LLMs for the generation of AI content, as well as for the detection, including ALLaM Bari et al.

(2024), Jais Sengupta et al. (2023), LLaMA-3.1 Grattafiori et al. (2024), and GPT-4 OpenAI (2023a). They also finetuned the XLM-RoBERTa Conneau et al. (2020) model to act as detectors, achieving notable results in identifying pure human content from artificial content. However, we did not include their detector in our experiments because according to Gillham, the model built by Originality.AI performs better and achieves state-of-the-art performance, Therefore, the results from Originality.AI are sufficient for our experiments. Additionally, we are not focusing on comparing every available detector. Instead, we are trying to determine whether slight polishing will affect the decision boundary of leading detectors, and it was very evident from the results of our experiments with Originality.AI Gillham and ZeroGPT Baroud.

3 Datasets

At the beginning, two datasets were constructed by collecting samples from popular news articles and blogs. The sources of the articles are listed in table 1. Three criteria were taken into account. First, we ensure that the articles collected are published in 2020 or before, so that they are ahead of the widespread use of LLMs. Second, to present different types of articles, we have collected articles from eight different domains, each contains 50 articles. Third, approximately half of the dataset was collected as news articles reporting something new or significant. The other half of the articles are simply general articles that discuss information in their field, not necessarily new.

Table 1 Sources of human-written articles used to construct the two datasets

Domain	Sources
Sport	Altamimi (2023) Einea (2019) Melhaj (2015) IJSSA
Tourisms	Altamimi (2023) Einea (2019) Wikipedia TheDollzter
Science	Melhaj (2015) Waqfeya UN Wikipedia
Technology	Altamimi (2023) Einea (2019) Melhaj (2015) Al-Awsat (2020)
Business	Coins4Arab (2020) Melhaj (2015)
Politics	Altamimi (2023) Einea (2019)
Medical	Altamimi (2023) Einea (2019) Melhaj (2015)
Finance	Einea (2019) Coins4Arab (2020)

For the first dataset, in addition to 400 articles that were written by humans, 10 LLMs were used to generate 400 additional AI articles. The LLMs that were used were (GPT-4o OpenAI (2024), GPT-3.5 OpenAI (2023b), LLaMA-3 70B Meta (2024), LLaMA-4 17B Meta (2025), Deepseek 3.1 Deepseek (2025), Claude-4 sonnet Anthropic (2024), Gemma-3 DeepMind (2025), Qwen-3 Alibaba (2025), Mistral Mensch (2024), and Kimi K2 Moonshot (2025)). Later on, we will use this dataset to assess how well LLMs and commercial AI models can distinguish AI-generated articles from human-generated articles.

Table 2 Comparison Between our Arabic Ar-APT dataset and the english APT dataset

Dataset	Language	# LLMs	# Domains	# Polishing Degrees	# Samples
Ar-APT	Arabic	10	8	4	16400
APT	English	5	6	11	15280

As for the second dataset, named Ar-APT, it uses the same 400 articles that were written by humans. Following that, we polished the articles at different percentage levels (10%, 25%, 50%, and 75%), We maintained only those four levels and did not include any more levels due to the fact that in this work Saha and Feizi (2025), which is similar to our work but done in English. There were 11 levels and results showed that AI-text detectors have difficulty in discriminating between minor and major polishing. As a result, adding a greater degree of AI polishing is not always associated with a higher detection rate. Furthermore, as compared to their work, instead of increasing the number of polishing levels, we have increased the number of LLMs used to polish. They polish their English articles by 5 LLMs, whereas in our work, we polish our Arabic articles by 10 LLMs. Table 2 compares our dataset with theirs. They have added more polishing levels, but we have dominated all other features, which are more crucial to our study. Ar-APT aims to determine whether slightly polished human-written articles will affect the borderline decision of LLMs and commercial AI models with regard to considering these articles to be AI-generated.

Following the collection of the second dataset, we applied two metrics to check whether polishing had been done correctly. For the purpose of determining if human-written text and AI-polished text retained the same meaning, we calculated the cosine similarity score between them. We have also done calculated Jaccard similarly score to verify that some change has actually been made according to the given polishing percentage. In Saha and Feizi (2025) they have removed samples that have a cosine similarity of less than 85%. However, in our case, the only samples that are excluded are those that are below 5%. Tables 4 and 5 illustrates each model along with the level of polishing and the resulting cosine similarity and Jaccard similarity scores.

In summary of this section, two datasets were created, one with 800 articles, half of which are human-written and half of which are AI-generated, and another dataset, Ar-APT, with 400 human-written artifices polished by 10 LLMs under four polishing percentages, resulting in a total of 16400 samples.

4 Experiments

4.1 Experiment setups

When it comes to AI text detectors, there has not been much research conducted in the arabic language. We were looking for models that are available to the general public users and can be used as AI text detectors. Hence, we used the same LLMs used for polishing the articles as AI detectors by only changing the System role prompt to "You are a helpful assistant who always determines if Arabic text is generated by AI

or by a Human. Only respond with AI or Human". Additionally, along with the LLMs, we have tested four commercial models, which are Oraginility.AI Gillham, ZeroGPT Baroud, Smodin Kevin, and Isgen Isgen.ai.

4.2 Evaluation metrics

There were two metrics used to measure the performance of the detectors: Accuracy, and False Positive Rate (FPR). Accuracy represents the percentage of articles that are not misclassified by the detectors. This is a useful method for determining the general robustness of the model. On the other hand, we use FPR to evaluate the percentage of articles authored by humans that have been misclassified as articles authored by AI. We are more concerned about avoiding falsely accusing the authors of AI plagiarism. Therefore, the FPR will assist us in protecting human authorship and measuring the trustworthiness of the detectors.

4.3 Experiments Results

In our experiments, we want to answer the following research questions (**RQs**):

RQ1: How effective are current LLMs and commercial AI detectors at detecting articles generated by artificial intelligence? **RQ2**: To what extent are LLMs capable of polishing Arabic text while maintaining the meaning? **RQ3**: Will polishing the article slightly or extensively affect LLMs decision as to whether the article was created by artificial intelligence? **RQ4**: Will polishing the article slightly or extensively affect the decision of the commercial AI detector to determine whether or not this article was generated by artificial intelligence?

4.3.1 Answering RQ1

In response to the first question, **RQ1**, do current LLMs and commercial AI detectors perform well in detecting AI-generated articles? We have tested 10 LLMs and 4 commercial AI detector models. Models that perform well will be nominated for answering **RQ3** and **RQ4**. According to table 3, there are four different types of metrics. The first metric (Original) measures the performance of the model in identifying human-written (original) text as it is. Similarly, the AI metric shows the performance of the model in classifying AI as AI. Third, (accuracy) measures the overall performance of the model in classifying both original and AI articles. Lastly, (FPR), which measures how often original articles is misclassified as AI.

The nominated LLMs are GPT-4o OpenAI (2024), Deepseek 3.1 Deepseek (2025), Mistral Mensch (2024), Claude-4 Sonnet Anthropic (2024), kimi K2 Moonshot (2025), LLaMA-4 17B Meta (2025), and Gemma-3 DeepMind (2025), while the nominated commercial models are orignality.AI Gillham and ZeroGPT Baroud. We chose these models for the following reasons. The FPR of GPT-4o, Deepseek 3.1, and Mistral is 6%, 3.13%, and 10.24%, respectively. This provides a greater level of assurance that human author articles will not be classified as AI. Meanwhile, they unfortunately classify the majority of AI as human-written (original), achieving an overall accuracy of 57.35%, 55.44%, and 51.61%, respectively. That is to say, they should be evolved in this regard in order to be used as AI detectors. The 3 LLMs strictly adhere to

Table 3 The accuracy of LLMs and commercial models in detecting AI-generated content. Original refers to evaluating the models on only human-authored articles. AI refers to evaluating the models on only AI-generated articles. Accuracy measure the overall performance across human-authored and AI-generated articles. FPR measure the model's ability to not label human-authored articles as AI.

Model	Original	AI	Accuracy	FPR
Large Language Models (LLMs)				
GPT-4	99.2	15.5	57.35	0.8
Deepseek 3.1	96.87	14.0	55.44	3.13
Mistral	89.76	13.46	51.61	10.24
Claude-4 Sonnet	83.51	83.17	83.34	16.49
LLaMA-4 17B	74.61	11.03	42.82	25.39
Kimi K2	74.20	75.94	75.07	25.80
Gemma-3-27B	52.74	63.00	57.87	47.26
Qwen-3	22.50	54.00	38.25	77.50
GPT-3.5	6.00	79.00	42.50	94.00
LLaMA-3 70B	11.00	86.25	48.63	89.00
Commercial Detectors				
Originality.AI	92.00	100.00	96.00	8.00
ZeroGPT	62.00	98.00	80.00	38.00
Isgen	43.00	53.00	48.00	57.00
Smodin	20.00	90.00	55.00	80.00

the principle of not falsely accusing original articles as AI-generated. Therefore, these LLMs are being nominated to serve as AI detectors to determine whether minor or major polishing will affect their decision and increase the FPR percentage, or whether they will continue to adhere to not falsely labeling original articles as AI.

In the case of Claude-4 Sonnet, which received 16.49% FPR, and Kimi K2, which received 25.8% FPR. Both models demonstrate good performance and acceptable FPR. Additionally, Claude-4 Sonnet and Kimi K2 achieved the highest overall accuracy, with 83.34% percent and 75.07%, respectively. These results indicate that these models are the most accurate among the 10 LLMs. Specifically, Claude-4 Sonnet achieved the highest level of accuracy as well as a low FPR. Thus, we have nominated them to act as AI detectors in the following experiments. The last two candidates to act as AI detectors in the following experiments are LLaMA-4 17B and Gemma-3. We consider these two even though they do not have a very low FPR as GPT-4o, Deepseek 3.1, and Mistral, and have not achieved a good performance compared to Claude-4 Sonnet and Kimi K2, they are included because they achieve an FPR lower than 50%. Including them will allow us to compare their results to those of the previous LLMs, in order to determine if polishing affects all models equally or if it has a more significant impact on models with lower performance. The three last models are Qwen-3, GPT 3.5, and LLaMA-3 70B. Achieve high FPRs of 77.5%, 94%, and 89%, respectively. Since they mostly misclassify human-authored articles as AI, they

cannot be trusted to act as AI detectors. Therefore, these three models have not been nominated for the second set of experiments.

Regarding commercial AI detectors models. It was important to include them because many companies and educational sectors rely on them to verify the authenticity of their articles. In addition, previous LLMs did not develop with the intention of serving as AI article detectors, whereas these commercial AI models do. There are two things we want to make sure: that they are trustworthy and that polishing will not affect their decisions. The first nominated Originality.AI Gillham has been selected as it obtained an impressive 8% FPR and 96% overall accuracy, which makes it the best model and very useful for our subsequent experiments to determine whether or not slightly polishing affects their results. ZeroGPT Baroud is the second nominated product, with a FPR of 38% and an overall accuracy of 80%. We have not nominated the last two sites, Smodin Kevin and Isgen Isgen.ai, as their results were not sufficient for the following experiments.

Additionally, our data contains 400 human-authored articles. It is important to note that when testing the performance of commercial models, we only test 100 articles and their polished versions. It is due to the fact that we were not able to obtain an API key in order to automate the process of testing articles. The process is carried out manually one by one. The two nominated models were tested on 100 original articles, and since each article was then polished 40 times with 10 models and under 4 different polishing settings, A total of 8000 articles were manually tested. Due to this, we did not test the entire dataset since we would need to manually test 32000 articles. Further, our main objective is to demonstrate that these commercial AI models will be affected by slight polishing, and our analysis of 100 articles is sufficient to demonstrate that.

4.3.2 Answering RQ2

Table 4 Average cosine similarity scores between original (human-authored) and AI-polished articles at different polishing levels

Model	10%	25%	50%	75%
GPT-4o	94.16	93.14	92.92	91.26
Deepseek 3.1	94.65	93.58	92.50	90.34
Mistral	94.82	94.44	92.70	92.05
Claude-4 Sonnet	95.75	94.62	89.34	89.16
LLaMA-4	93.28	91.30	90.51	88.20
Gemma-3	93.12	93.15	92.54	90.54
Qwen-3	80.94	82.20	81.66	81.46
GPT-3.5	92.66	92.63	92.67	92.28
LLaMA-3 70B	92.25	88.95	88.06	85.40
Kimi K2	92.26	88.28	89.39	88.36

The table 4 shows that LLMs polish the data and retain the meaning to such an extent that all LLMs, with the exception of Qwen-3, Alibaba (2025), achieve an

Table 5 Average Jaccard similarity scores between original (human-authored) and AI-polished articles at different polishing levels

Model	10%	25%	50%	75%
GPT-4	66.68	47.50	41.21	29.01
DeepSeek (DEEPS)	70.03	58.71	45.39	28.50
Mistral (MIST)	64.85	50.67	45.17	36.81
Claude-4	89.15	80.00	35.93	34.82
LLaMA-4	64.90	54.79	45.01	35.17
Gemma-3-27B	45.38	41.15	40.63	26.40
Qwen	36.45	27.73	30.82	19.75
GPT-3.5	73.81	73.59	73.72	73.62
LLaMA-3-70B	60.96	48.48	42.99	32.55
Kimi	32.22	23.24	23.97	20.81

average cosine similarity score above 92% under 10% polishing settings. Increasing the polishing percentage reduces this cosine similarity, but not by a substantial amount. LLaMA-3 70B is the only model that decreases dramatically, Meta (2024), achieving a good cosine similarity of 92.25% at a polishing setting of 10% and drops to 85.4% at a polishing setting of 75%.

The best model for preserving the meaning between them is Mistral Mensch (2024), which achieves an average of 93.5% across all polishing settings. Then followed by GPT-4o OpenAI (2024), and Deepseek 3.1 Deepseek (2025), each of which achieved an average of 92.87% and 92.77% respectively. The worst model is Qwen-3, which achieves an average of 81.57%. This indicates that Qwen-3 is not suitable for polishing Arabic texts.

A cosine similarity score of less than 100% between two sentences does not necessarily mean that they do not have the exact meaning; rather, it indicates that in the embedding space, the sentences do not share the exact placement of words. For example, the two sentences "He began working this morning." and "He began working today morning.". These sentences have exactly the same meaning in the reader's mind, but the cosine similarity score is 96.49% with SBERT Reimers and Gurevych (2020) embedding space and 97.64% with MPNet Song et al. (2020) embedding space. Accordingly, a polished article having more than 90% cosine similarity score indicates that it preserves the original meaning. All used LLMs except Qwen-3 are considered to be good models for polishing Arabic text while maintaining its meaning.

From table 5, we can see similar ideas to table 4. The Jaccard similarity score is highest at the 10% polishing setting and decreases as the polishing percentage is increased. Moreover, we notice that it does not adhere strictly to the percentage of polishing given to them. At a polishing setting of 10%. The best model is Claude-4 Sonnet Anthropic (2024), which achieved 89.15%. After that, GPT 3.5 OpenAI (2023b), Deepseek 3.1 Deepseek (2025), and GPT-4o OpenAI (2024) respectively achieved 73.81%, 70.03%, and 66.68%. It is acceptable to consider this as a minor change as

long as the Jaccard similarity score is above 60% in our experiment. Therefore, Qwen-3 Alibaba (2025), Kimi K2 Moonshot (2025) and Gemma-3 DeepMind (2025) cannot be considered as having done minor polishing under 10% and 25% polishing settings.

To demonstrate how the Jaccard similarity calculation is performed, we will use these two sentences, the first sentence of 10 words. "we went for the grocery store to buy some milk". The second sentence also contains 10 words. "we went for the grocery store to get some milk". Despite the fact that they both contain exactly the same number of words and the same words with one exception. Since the Jaccard calculation formula is (Total intersect words/Total unique words in both sentences), the Jaccard similarity score between these two sentences will be 81.82%. As a measure of polishing quality, it is helpful to use Jaccard to verify that the text has not undergone too many changes.

As a final conclusion from these two tables. Answering **RQ2**, all 10 LLMs are capable of polishing arabic text, but their performance differs from one another, with Claude-4 Sonnet being the best followed by Deepseek 3.1 and GPT-4o being the next-best. Therefore, for **RQ3** experiments, we should pay more attention to Claude-4 Sonnet results. since it attains a high cosine similarity score and follows a good polishing percentage. Afterwards, Deepseek 3.1 and GPT-4o both achieved good results, but were not on the same level as Claude-4 Sonnet. Follow that Mistral Mensch (2024), LLaMA-3 70B Meta (2024) and LLaMA-4 17B Meta (2025). They all achieved good results, but not at the same level as the above-mentioned models. The cosine similarity and Jaccard similarity scores for GPT-3.5 OpenAI (2023b) were similar in both tables, indicating that it does not listen to the prompt instructions. The situation of GPT-3.5 has been reported in several publications such as Almohaimeed (2024b) that it is not that good at following Arabic instructions. The Jaccard similarity scores of Qwen-3, Kimi K2, and Gemma-3 are very low, under 10% and 25% polishing settings. Thus, when answering **RQ3**, if these three models consider the article to be AI-generated, it will not be taken into account.

4.3.3 Answering RQ3

In this experiment, our goal is to demonstrate that we do not want to falsely accuse writers that their articles are AI-generated. As they only use current LLMs to improve their own writing without changing the meaning or changing too much of their own style. Detailed results of the experiments are presented in figures [2,3,4,5,6,7] and as numerical representations in tables [A1, A2, A3 A4] in Appendix A.

From our answer to **RQ1**, we have stated that the GPT-4o, Deepseek 3.1, and Mistral models strictly adhere to the principle of not falsely accusing original articles of being generated by artificial intelligence. In the case of GPT-4o OpenAI (2024) detectors, which achieved 99.2% under 0% polishing, the performance of detecting did not significantly change under 10% polishing, with the exception of articles polished by LLaMA-3 70B, which achieved 95.26%. Approximately 5% of articles are falsely accused of being AI-generated. Under 25% polishing setting, detecting articles polished by Mistral is 97.30%, LLaMA-4 17B is 96.30%, and LLaMA-3 70B is 87.04%. This illustrates that even a robust model such as GPT-4o can become affected by minor or mid-minor polishing. Additionally, GPT-4o performed well and did not accuse other

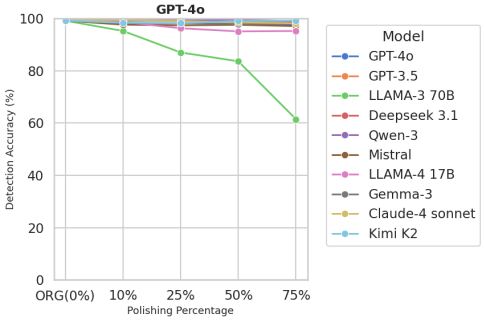

Fig. 2 AI-text detection rate for GPT-4o with 5 different polishing settings.

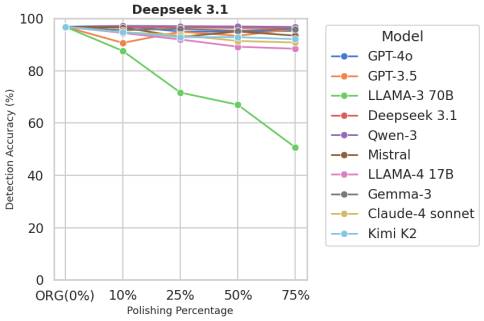

Fig. 3 AI-text detection rate for Deepseek 3.1 with 5 different polishing settings.

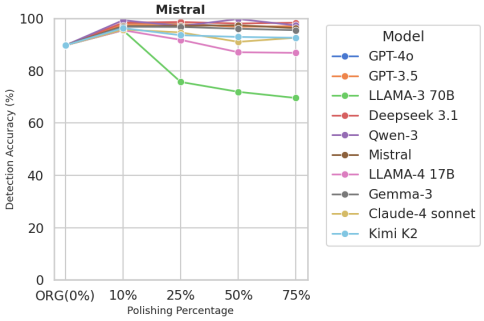

Fig. 4 AI-text detection rate for Mistral with 5 different polishing settings.

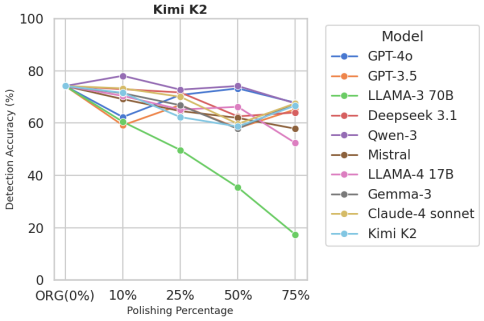

Fig. 5 AI-text detection rate for Kimi K2 with 5 different polishing settings.

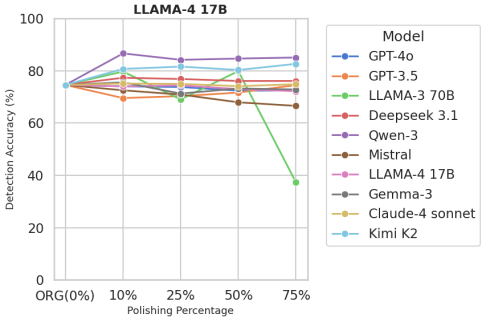

Fig. 6 AI-text detection rate for LLaMA-4 17B with 5 different polishing settings.

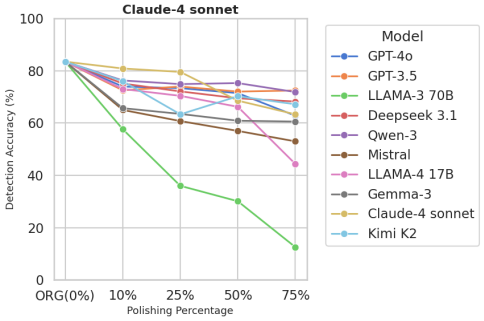

Fig. 7 AI-text detection rate for Claude-4 Sonnet with 5 different polishing settings.

articles polished by Deepseek 3.1, Qwen-3, and Claude-4 Sonnet, having all achieved greater than 99% under 10% and 25% polishing settings. The OpenAI OpenAI platform offers researchers a way to partially finetune their models. Future researchers

may utilize this functionality to partially fine-tune GPT-4o and make it differentiate between AI-generated and slightly polished human-authored articles.

Deepseek 3.1 Deepseek (2025), which received 3.13 FPR, was expected to not be affected too much by slightly polishing articles, but that was not the case. When the articles were polished by LLaMA-3 70B, deepseek 3.1 obtained a 9.24% lower percentage under 10% polishing settings and a 24.91% lower percentage under 25% polishing settings. Even though it was not affected by the polishing of other models as happened with LLaMA-3 70B, it is a red flag that we need to consider in order to avoid false accusations. The same problem also occurs with Mistral Mensch (2024), which performs well across all articles generated by other models except LLaMA-3 70B. There is no doubt that the LLaMA-3 70B polishing style structure tricks the model into thinking that it is AI-generated. Even under 50% and 75% polishing settings, GPT-4o, Deepseek 3.1 and Mistral decrease slightly in all other nine models with the exception of LLaMA-3 70B. This illustrates that LLaMA-3 70B approach in handling Arabic words differs from all other models. The effect of LLaMA-4 17B is somewhat similar to that of LLaMA-3 70B, however, it is not by as much.

Furthermore, The best performing LLM, Claude-4 Sonnet Anthropic (2024), when evaluated under 10% polishing settings, we can observe that it has become the most affected LLM. In the case of 0% polishing, the level of accuracy was 83.51%. Under 10% polishing, the level of accuracy decreased significantly for all LLMs, where the lowest decrease was obtained with articles polished by their own, with a decrease of 3%, and the highest decrease was obtained with articles polished by LLaMA-3 70B, achieving approximately 26% lower accuracy. Under 25% polishing, results decreased by more than 20% with models such as Kimi K2, Gemma-3, and Mistral. Additionally, performance of Claude-4 Sonnet with articles polished LLaMA-3 70B experienced a significant reduction by almost 50%. The significant reduction in accuracy emphasizes the need for building an AI detector that can distinguish between articles created by AI and articles that have been slightly polished by humans.

LLaMA-4 17B Meta (2025) and Kimi K2 Moonshot (2025) did not achieve that good results. Their accuracy was 75.07% and 42.82%, respectively, and their FPR was 25.8% and 25.39%. The purpose of reporting the results of these two detectors is to illustrate the difference between their performance and that of better LLM detectors, including GPT-4o, Deepseek 3.1, Mistral, and Claude-4 Sonnet. Both of them obtained similar results. However, Kimi K2 has a lower FPR rate when increasing the polishing percentage, but this is not surprising given that the model is not constructed to function as an AI detector.

4.3.4 Answering RQ4

According to table 3, Originality.AI Gillham has achieved a total accuracy of 96% and a false positive rate of 8%. When it is tested under four different polishing settings. We can observe how badly it is affected under even the smallest amount of change. The performance of Originality.AI on articles polished by GPT-4o decreases by 69%. It is likely that these models are built to prove themselves to be effective AI detectors, and since GPT-4o is currently the most commonly used model. The Originality.AI tries to demonstrate its abilities by training on large amounts of data collected from

Table 6 AI-text detection accuracy for Originality.AI commercial model under 4 different polishing settings, Model refers to the ones used for polished the text

Model	10%	25%	50%	75%
GPT-4o	23	23	12	0
GPT-3.5	74	74	74	74
Kimi K2	28	28	25	35
Claude-4 Sonnet	90	90	74	71
LLaMA-3 70B	76	76	46	44
deepseek 3.1	60	60	32	21
Qwen-3	18	18	18	0
Mistral	12	12	9	7
LLaMA-4 17B	71	71	37	35
Gemma-3	12	12	12	5

Table 7 AI-text detection accuracy for ZeroGPT commercial model under 4 different polishing settings, Model refers to the ones used for polished the text

Model	10%	25%	50%	75%
GPT-4o	33	33	31	29
GPT-3.5	53	51	54	53
Kimi K2	50	48	47	45
Claude-4 Sonnet	57	57	57	57
LLaMA-3 70B	55	53	53	57
deepseek 3.1	54	53	51	50
Qwen-3	42	42	42	41
Mistral	31	31	31	31
LLaMA-4 17B	43	43	43	47
Gemma-3	34	34	33	33

such a model. However, for models like Claude-4 Sonnet under 10% and 25% polishing settings, the results of the Originality.AI detection are not dramatically different. This is likely because the data used to train the Originality.AI model does not have the AI article style generated by Claude-4 Sonnet. Moreover, for articles polished using Mistral or Gemma-3, the Originality.AI model falsely accuses all articles except for 12%.

This calls for researchers and commercial AI developers to construct models that can differentiate between AI-generated articles and original articles that have been slightly polished by AI. The problem of falsely accusing authors of utilizing an AI-generated tool to generate their text is more significant than failing to address actual violators. As for the results related to the ZeroGPT Baroud, even though it did not achieve very high results in Arabic as Originality.AI, ZeroGPT achieved 80% total accuracy and 38% FPR. Under 10% polishing settings, the highest decrease in results

was for articles polished by Mistral, GPT-4o, and Gemma-3, which amounted to 31%, 33%, and 34%, respectively.

In table 6, We can observe that GPT-4o and Claude-4 Sonnet experiences a sharp decline under the 4 polishing setting. This is due to the level of Jaccard similarity differences observed in GPT-4o and Claude-4 Sonnet in table 5. As opposed to GPT 3.5, which is very stable across all polishing settings. It is due to the same reason that the Jaccard and cosine similarities scores remained the same under all four polishing settings in tables 4 and 5. For results related to ZeroGPT in table 7, In the case of Claude-4 Sonnet, we observe the same results 57% regardless of the polishing setting, even though the Jaccard similarity varies among the polishing settings. It is probably because models such as ZeroGPT are trained to detect articles generated by GPT-4o and LLaMA-3 as they stated in their official website, but did not pay much attention to models such as Claude-4 Sonnet, which are less commonly used. We have also noticed that for many models, such as the LLaMA-3 70B, Deepseek 3.1, Qwen-3, there is no significant difference when polishing under 10% or 75%. This could be for the same reason mentioned with articles polished by Claude-4 Sonnet.

5 Conclusion

This paper presents a comprehensive analysis of the performance of 10 popular large language models such as GPT-4o, LLaMA-3, and Claude-4 Sonnet, as well as 4 commercial AI detection such as Originality.AI and ZeroGPT, in order to address two primary concerns. First concern focuses on evaluating the ability of those models to distinguish between Arabic articles authored by humans and Arabic articles generated by AI. To assess this concern, we have constructed a dataset consisting of 800 samples. The best model results were obtained with the Claude-4 Sonnets, which had an overall accuracy of 83.51% and a FPR of 16.49%. Additionally, Originality.AI is the best commercial AI model, with an accuracy rate of 96% and a FPR of 8%. On the other side, our second concern focuses on evaluating the same models to determine if using AI to slightly polish a human-written article without changing its meaning will result in changing the decision of detectors from a human-written article to an entirely AI-generated article. It is intended to assess the sensitivity of these detectors to slight polishing applied to human-written text. An Ar-APT dataset of 16400 samples has been constructed to assess this concern, and the results indicate that the best performing LLM, Claude-4 Sonnet, was adversely affected across all articles polished by any of the 10 LLMs. Articles generated by LLaMA-3 70B demonstrated the highest performance decrease under 10% polishing settings, with Claude-4 Sonnet performance decreasing from 83.51% to 57.63%. On the other side, Commercial AI models were more affected than LLMs as the performance of Originality.AI models declined by 80% from 92% to 12% with articles polished by Mistral or Gemma-3 under 10% polishing setting. To avoid falsely accusing authors of AI palagrisms and to increase the trustworthiness of AI detectors, researchers and AI detector developers should pay more attention to consider slightly polished articles as not AI generated.

Acknowledgments This research was funded by the Ongoing Research Funding program, (ORF-2025-XXX), King Saud University, Riyadh, Saudi Arabia.

Author contributions: Saleh Almohaimeed: Conceptualization, Study Design, Dataset Preparation, Experimentation, Result Analysis, Writing Original Draft, Writing Review and Editing. Saad Almohaimeed: Conceptualization, Comparative Experiments, Model Evaluation, Writing Review and Editing. Mousa Jari: Conceptualization, Writing Review and Editing. Khaled A. Alobaid: Conceptualization, Writing Review and Editing. Fahad Alotaibi: Conceptualization, Writing—Review and Editing.

Data availability: The data in this study are available at https://github.com/Saleh-Almohaimeed/Ar-APT

This is a preprint version of the manuscript.

References

Asharq Al-Awsat. Asharq al-awsat: The leading pan-arab international newspaper, 2020. URL https://aawsat.com. Accessed 2025-11-08.

Maged S. Al-Shaibani and Moataz Ahmed. The arabic ai fingerprint: Stylometric analysis and detection of large language models text. *arXiv preprint arXiv:2505.23276*, 2025.

Alibaba. Qwen-3: Next-generation multilingual foundation model. *Alibaba Cloud Research Report*, 2025. URL https://qwenlm.github.io/.

Saad Almohaimeed. Thos: A benchmark dataset for targeted hate and offensive speech. In *Proceedings of the Data-centric Machine Learning Research Workshop at ICML 2023*, 2023. arXiv preprint arXiv:2311.06446.

Saad Almohaimeed. Transfer learning and lexicon-based approaches for implicit hate speech detection: A comparative study of human and gpt-4 annotation. In *Proceedings of the 18th IEEE International Conference on Semantic Computing (ICSC-2024)*, pages 142–147, 2024a.

Saleh Almohaimeed. GAT-SQL: an advanced prompt engineering approach for effective text-to-sql interactions. In *2024 IEEE Congress on Evolutionary Computation (CEC)*, pages xxx–xxx, 2024b. doi: 10.1109/CEC60901.2024.xxxxxx.

Hamed Alshammari, Ahmed El-Sayed, and Khaled Elleithy. Ai-generated text detector for arabic language using encoder-based transformer architecture. *Big Data and Cognitive Computing*, 8(3):32, 2024. doi: 10.3390/bdcc8030032.

Mohammed Altamimi. ANAD: arabic news article dataset. *Data in Brief*, 50:109460, 2023. doi: 10.1016/j.dib.2023.109460.

Anthropic. Claude 4 sonnet: Advancements in scalable constitutional ai. *Anthropic Research Report*, 2024.

M. Saiful Bari, Yazeed Alnumay, Norah A. Alzahrani, Nouf M. Alotaibi, Hisham A. Alyahya, Sultan AlRashed, Faisal A. Mirza, Shaykhah Z. Alsubaie, Hassan A. Alahmed, Ghadah Alabduljabbar, and .... ALLaM: large language models for arabic and english. *arXiv preprint arXiv:2407.15390*, 2024.

Rawad Baroud. Zerogpt: Ai content detection tool for gpt-generated text. *ZeroGPT Research Blog*. URL https://www.zerogpt.com.

Chaka Chaka. Reviewing the performance of AI detection tools in differentiating between AI-generated and human-written texts: A literature and integrative hybrid review. *Journal of Applied Learning & Teaching*, 7(1):115–126, 2024. doi: 10.37074/jalt.2024.7.1.14.

Coins4Arab. Coins4arab – arabic cryptocurrency forum, 2020. URL https://coins4arab.com/vb/index.php. Accessed 2025-11-08.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020)*, pages 8440–8451. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.747.

Google DeepMind. Gemma 3: Open models for responsible ai innovation. *Google DeepMind Technical Report*, 2025. URL https://deepmind.google/technologies/gemma/.

Deepseek. Deepseek-v3.1: Advanced multilingual large language model. *DeepSeek AI Technical Report*, 2025. URL https://www.deepseek.com/.

Liam Dugan, Alyssa Hwang, Filip Trhlík, Josh Magnus Ludan, Andrew Zhu, Hainiu Xu, Daphne Ippolito, and Chris Callison-Burch. Raid: A shared benchmark for robust evaluation of machine-generated text detectors. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024)*, pages 12463–12492, 2024.

Omar Einea. SANAD: single-label arabic news articles dataset for automatic text categorization. *Data in Brief*, 25:104 076, 2019. doi: 10.1016/j.dib.2019.104076.

Colleen Farrelly and Singh. Current topological and machine learning applications for bias detection in text. In *2023 6th International Conference on Signal Processing and Information Security (ICSPIS)*, pages 190–195, 2023. doi: 10.1109/ICSPIS60075.2023.10343824.

Sebastian Gehrmann, Hendrik Strobelt, and Alexander M. Rush. GLTR: statistical detection and visualization of generated text. *arXiv preprint arXiv:1906.04043*, 2019.

Jon Gillham. Originality.ai: Advanced ai content detection and plagiarism analysis platform. *Originality.AI Research Blog*. URL https://originality.ai.

Jonathan Gillham. Study finds almost 11 *Originality.ai Blog*, 2025.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, and .... The LLaMA 3 Herd of

Models. *arXiv preprint arXiv:2407.21783*, 2024.

IJSSA. The scientific journal of sport science & arts. URL https://ijssa.journals.ekb.eg/journal/about?lang=en. Accessed 2025-11-08.

Isgen.ai. Isgen ai detector – detect ai-generated text (arabic & multilingual). URL https://isgen.ai/ar. Accessed 2025-11-09.

Kevin. Smodin ai detector – accurate ai checker for chatgpt, gpt-5 & gemini. URL https://smodin.io/ai-content-detector. Accessed 2025-11-09.

Yafu Li, Qintong Li, Leyang Cui, Wei Bi, Zhilin Wang, Longyue Wang, Linyi Yang, Shuming Shi, and Yue Zhang. MAGE: machine-generated text detection in the wild. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024)*, pages xxx–xxx, 2024. URL https://aclanthology.org/2024.acl-long.3.pdf. Dataset available at https://github.com/yafuly/MAGE.

Dominik Macko, Robert Moro, Adaku Uchendu, Jason S. Lucas, Michiharu Yamashita, Matúš Pikuliak, Ivan Srba, Thai Le, Dongwon Lee, Jakub Simko, and Maria Bielikova. MULTITuDE: Large-scale multilingual machine-generated text detection benchmark. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP 2023)*, 2023.

Mohammed Melhaj. EASC: essex arabic summaries corpus. Corpus release, 153 articles + 765 summaries, 2015. Available at SourceForge.

Arthur Mensch. Mistral-seba 24: Efficient and open-weight large language model. *Mistral AI Research Report*, 2024. URL https://mistral.ai/news/mistral-7b/.

Meta. The llama 3 herd of models. *Meta AI Research Report*, 2024.

Meta. LLaMA-4 Maverick-17B-128E-Instruct-FP8. https://huggingface.co/meta-LLaMA/LLaMA-4-Maverick-17B-128E-Instruct-FP8, 2025. Mixture-of-Experts (MoE) model: 17B active params / 128 experts / FP8 quantization. Released April 2025. Knowledge cutoff Aug 2024. :contentReference[oaicite:2]index=2.

Moonshot. Kimi-K2-Instruct: A 1 Trillion-Parameter Mixture-of-Experts Language Model (32 B active) for Agentic Intelligence. https://huggingface.co/moonshotai/Kimi-K2-Instruct, 2025. Released July 2025; supports 128 K context window. :contentReference[oaicite:2]index=2.

OpenAI. Openai platform – developer api & resources. URL https://platform.openai.com/. Accessed 2025-11-09.

OpenAI. GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774*, 2023a.

OpenAI. GPT-3.5-Turbo: Advancements in Instruction-Tuned Large Language Models. *OpenAI Technical Report*, 2023b. URL https://platform.openai.com/docs/models/gpt-3-5-turbo. Accessed 2025-11-08.

OpenAI. Gpt-4o: Omnimodal large language model. *OpenAI Technical Report*, 2024.

Shaina Raza, Muskan Garg, Deepak John Reji, Syed Raza Bashir, and Chen Ding. Nbias: A natural language processing framework for bias identification in text. *Expert Systems With Applications*, 237:121542, 2023. doi: 10.1016/j.eswa.2023.121542.

Nils Reimers and Iryna Gurevych. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, page in press, 2020.

Shoumik Saha and Soheil Feizi. Almost ai, almost human: The challenge of detecting ai-polished writing. *Findings of the Association for Computational Linguistics: ACL 2025*, pages 25414–25431, 2025.

Areg Mikael Sarvazyan, José Ángel González, Marc Franco-Salvador, Francisco Rangel, Berta Chulvi, and Paolo Rosso. Overview of AuTexTification at IberLEF 2023: Detection and Attribution of Machine-Generated Text in Multiple Domains. *Procesamiento del Lenguaje Natural*, 71:275–288, 2023.

Nandika Sengupta, Sudip Kumar Sahu, Bo Jia, Sriram Katipomu, Hu Li, Felix Koto, Will Marshall, Gem Gosal, Chen Liu, Zhen Chen, and .... Jais and Jais-Chat: Arabic-centric foundation and instruction-tuned open generative large language models. *arXiv preprint arXiv:2308.16149*, 2023.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. MPNet: masked and permuted pre-training for language understanding. In *Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS 2020)*, pages 16857–16867. Curran Associates, Inc., 2020. doi: 10.5555/3495724.3497138.

Zhenpeng Su. Hc3 plus: A semantic-invariant human chatgpt comparison corpus. Dataset release ..., 2024. Supports wide variety of languages.

TheDollzter. The dollzter travels – a travel and lifestyle blog featuring destinations, experiences, and guides around the world. URL https://www.thedollztertravels.com/. Accessed 2025-11-08.

UN. United nations – official website: Global organization for peace, security, and cooperation. URL https://www.un.org. Accessed 2025-11-08.

Hao Wang, Jianwei Li, and Zhengyu Li. Ai-generated text detection and classification based on bert deep learning algorithm. *arXiv preprint arXiv:2405.16422*, 2024a.

Yuxia Wang, Jonibek Mansurov, Petar Ivanov, Jinyan Su, Artem Shelmanov, Akim Tsvigun, Chenxi Whitehouse, Osama Mohamed Afzal, Tarek Mahmoud, Toru Sasaki, Thomas Arnold, Alham Fikri Aji, Nizar Habash, Iryna Gurevych, and Preslav Nakov. M4: Multi-generator, multi-domain, and multi-lingual black-box machine-generated text detection. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1369–1407. Association for Computational Linguistics, 2024b. doi: 10.18653/v1/2024.eacl-long.83. URL https://aclanthology.org/2024.eacl-long.83.

Waqfeya. Waqfeya - arabic digital library for free downloadable books and manuscripts. URL https://waqfeya.com/. Accessed 2025-11-08.

Wikipedia. Wikipedia: The free encyclopedia – a collaborative multilingual knowledge platform. URL https://ar.wikipedia.org. Accessed 2025-11-08.

Han Xu, Jie Ren, Pengfei He, Shenglai Zeng, Yingqian Cui, Amy Liu, Hui Liu, and Jiliang Tang. On the Generalization of Training-based ChatGPT Detection Methods. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7223–7243. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-emnlp.424. URL https://aclanthology.org/2024.findings-emnlp.424/.

Appendix A Numerical representation of figures 2 to 7

Table A1 AI-text detection accuracy of 6 LLMs, where columns refer to those used for detection, and rows refer to those used for polishing the text under 10% polishing. ORG refers to the accuracy of the detector under no polishing.

Models	**GPT-4o**	**Deepseek 3.1**	**Mistral**	**Kimi K2**	**LLAMA-4 17B**	**Claude-4 sonnet**
ORG	99.2	96.87	89.76	74.2	74.61	83.51
GPT-4o	99.2	97.12	98.17	62.3	74.08	73.99
GPT-3.5	97.88	90.7	97.88	59.15	69.58	72.54
LLAMA-3 70B	95.26	87.63	95.53	60.5	79.73	57.63
Deepseek 3.1	99.48	96.6	98.43	72.97	77.35	75.13
Qwen-3	99.7	97.23	99.38	78.1	86.67	76.31
Mistral	97.67	96.63	97.15	69.17	72.54	65.03
LLAMA-4 17B	98.69	94.5	95.55	70.42	74.08	73.04
GEMMA	98.22	95.58	96.89	71.43	75.58	65.71
Claude-4 sonnet	99.48	94.76	95.81	73.3	75.13	80.89
Kimi K2	98.43	94.79	96.35	71.61	80.73	76.04

Table A2 AI-text detection accuracy of 6 LLMs, where columns refer to those used for detection, and rows refer to those used for polishing the text under 25% polishing. ORG refers to the accuracy of the detector under no polishing.

Detectors	**GPT-4o**	**Deepseek 3.1**	**Mistral**	**Kimi K2**	**LLAMA-4 17B**	**Claude-4 sonnet**
ORG	99.2	96.87	89.76	74.2	74.61	83.51
GPT-4o	99.2	95.04	97.90	70.76	73.80	73.37
GPT-3.5	98.21	94.97	97.88	66.93	70.31	73.99
LLAMA-3 70B	87.04	71.69	75.75	49.7	69.05	35.98
Deepseek 3.1	99.29	96.72	98.70	71.73	76.88	72.13
Qwen-3	99.32	97.10	97.14	72.80	84.22	74.92
Mistral	97.43	93.00	97.22	64.61	70.92	60.77
LLAMA-4 17B	96.30	92.00	91.87	65.18	74.42	70.42
GEMMA	98.30	96.10	96.80	66.84	71.30	63.51
Claude-4 sonnet	99.06	93.67	94.71	70.08	75.01	79.58
Kimi K2	98.22	92.91	93.63	62.30	81.62	63.35

Table A3 AI-text detection accuracy of 6 LLMs, where columns refer to those used for detection, and rows refer to those used for polishing the text under 50% polishing. ORG refers to the accuracy of the detector under no polishing.

Detectors	**GPT-4o**	**Deepseek 3.1**	**Mistral**	**Kimi K2**	**LLAMA-4 17B**	**Claude-4 sonnet**
ORG	99.2	96.87	89.76	74.2	74.61	83.51
GPT-4o	98.43	95.03	97.01	73.28	72.41	71.47
GPT-3.5	98.22	93.42	96.89	58.05	71.68	72.05
LLAMA-3 70B	83.64	67.02	71.96	35.4	79.9	30.08
Deepseek 3.1	98.91	96.42	98.12	62.57	76.12	69.49
Qwen-3	99.64	96.99	99.83	74.2	84.70	75.31
Mistral	97.62	95.06	97.45	61.98	67.95	57.00
LLAMA-4 17B	95.10	89.23	87.12	66.23	73.04	66.23
GEMMA	97.90	95.33	96.12	57.81	73.25	60.90
Claude-4 sonnet	98.55	91.44	91.10	59.69	74.13	68.59
Kimi K2	99.12	92.88	92.98	58.68	80.33	70.26

Table A4 AI-text detection accuracy of 6 LLMs, where columns refer to those used for detection, and rows refer to those used for polishing the text under 75% polishing. ORG refers to the accuracy of the detector under no polishing.

Detectors	**GPT-4o**	**Deepseek 3.1**	**Mistral**	**Kimi K2**	**LLAMA-4 17B**	**Claude-4 sonnet**
ORG	99.2	96.87	89.76	74.2	74.61	83.51
GPT-4o	97.6	95.29	96.86	67.8	72.51	62.83
GPT-3.5	98.4	96.02	97.08	65.25	74.54	72.49
LLAMA-3 70B	61.5	50.67	69.6	17.3	37.33	12.53
Deepseek 3.1	98.95	96.33	98.43	64.04	76.12	68.24
Qwen-3	98.84	96.81	97.39	67.54	85.10	71.88
Mistral	97.13	93.47	96.34	57.85	66.58	53.00
LLAMA-4 17B	95.26	88.45	86.88	52.37	72.20	44.36
GEMMA	98.17	96.03	95.56	67.62	72.85	60.57
Claude-4 sonnet	98.17	90.84	92.67	67.54	74.87	63.35
Kimi K2	99.21	92.13	92.65	66.67	82.67	67.19

Appendix B Examples of System and User Prompts

System Prompt

```
You are a helpful assistant who always responds with helpful information.
You are asked to provide a polished version of the given text.
Only generate the Arabic polished text.
```

User Prompt

```
You are an editor that polishes Arabic text.
Your task is to edit a given Arabic article by changing a specific percentage
of words.
You must follow these rules:
- Keep the meaning and message of the text unchanged.
- Do not shorten or expand the length significantly.
- Replace/rephrase only the target percentage of words.
- Output only the polished Arabic text, nothing else.

### Now polish this new text by 75% only:
$Article
```

Fig. B1 This is an example of the prompt instruction used by the 10 LLMs to polish the human-written articles, where the percentage specified will be one of four: 10%, 25%, 50% and 75%.

System Prompt

```
You are a helpful assistant who always determines if Arabic text
 is generated by AI or by a Human. Only respond with AI or Human
```

User Prompt

```
Determine if the text is generated by AI or by a human
Note: Give the requested answer without any explanation.
$Article
```

Fig. B2 This is an example of the prompt instruction used by the 10 LLMs to detect AI-generated articles. we emphasize that LLM should respond by human or artificial intelligence, as we see that this emphasis will increase their compliance with this instruction