Received 6-09-2024; Accepted 6-11-2024; Published 13-11-2024

A Comparative Study in Large Language Models Usage for Fake News Detection

Repede Ştefan Emil Stefan.repede@ulbsibiu.ro
"Lucian Blaga" University, Sibiu, Romania,
Faculty of Engineering, Field of Computer Engineering and Information Technology,
Postal address: Bld. Victoriei, N.10, Sibiu, 550024, Romania.

Remus BRAD remus.brad@ulbsibiu.ro
"Lucian Blaga" University, Sibiu, Romania,
Faculty of Engineering, Field of Computer Engineering and Information Technology,
Postal address: Bld. Victoriei, N.10, Sibiu, 550024, Romania.

Corresponding Author: Remus BRAD

Abstract

This study presents a comparative analysis of the capabilities of Large Language Models (LLMs) in the automatic detection of fake news. The research focuses on evaluating the accuracy, precision, recall and F1-score of three state-of-the-art LLMs—OpenAI's ChatGPT 4.0 beta, Meta's Llama 3.1, and Google's Gemini—using an extensive dataset of verified true and fake news. By employing a black-box testing method, the study categorizes the LLMs' outputs and assesses their performance based on accuracy metrics. The results indicate moderate success in distinguishing between true and false news, with differences noted between the models and smaller, text classification specialized Natural Language Models (NLP), like Google's Bidirectional Encoder Representations from Transformers (BERT) model variants, trained on the same fake news dataset. The findings demonstrate the potential of LLMs as tools for combating misinformation, while also emphasizing the current limitations and the need for improvements in their accuracy and reliability. This paper provides insights into the challenges of utilizing LLMs for misinformation detection and highlights the importance of combining technological advancements with our distinct human cognition.

Keywords: ChatGPT, LLaMA 3.1, Fake news, Disinformation, Gemini

1. INTRODUCTION

due to their data analysis possibilities like summarization and information retrieval Large Language Models (LLMs) can prove useful in fact-checking tasks that are usually carried by human operators due to their significant complexity. As the new LLM models are gaining recognition as a highly reliable technology in various fields they are also anticipated to attract many users seeking to validate

Citation: Repede Ştefan Emil and Remus BRAD. A Comparative Study in Large Language Models Usage for Fake News Detection. Advances in Artificial Intelligence and Machine Learning. 2024;4(4):163.

different types of news. Despite their hype, conflicting initial studies on LLM model accuracy reveal that they perform exceptionally well on smaller sample sizes [1], but can generate unexpected and sometimes misleading results called “hallucinations”, either when given straightforward or complex prompts, highlighting their unpredictable nature [2].

The current leading LLMs in the English-speaking areas are Google’s Gemini, Meta‘s Llama 3.1 and OpenAI’s ChatGPT 4.0 beta. Generalizing about the performance of these LLMs in fake news detection can be potentially misleading due to the significant differences in the training data they are exposed to and the fact that all the models selected have access to the internet [3], which should grant them an extremely high percentage in accuracy while dealing with already available fact-checked data.

ChatGPT models up to version 4.0 have a significant drawback, which is their limited ability to access up-to-date information beyond a specific date [4]. The same drawback can be found on the LLaMA 2 (Large Language Model Meta AI) series from Meta which was released in July 2023 [5]. However, both for OpenAI’s ChatGPT and for Meta’s LLaMA 3.1 versions these limitations are absent, thus both models can browse the web and gather additional input [4]. In contrast, Gemini never had such restrictions and can access the web for their responses. However, it is interesting to note that when comparing their responses’ formatting, including aspects like length, topic coverage, accuracy, etc., they are evidently different. Additionally, ChatGPT’s and Gemini’s premium versions are not free and impose caps on the number of responses per period and further ethical limitations which can hinder their responsiveness when dealing with politics-based text [6].

To determine the optimal performance of the LLMs, multiple metrics may be used. However, this paper focuses on accuracy, precision, recall and F1-score metrics for fake news detection [7]. The three prevailing LLMs, (ChatGPT, Gemini and Llama) will be subjected to investigative testing to evaluate their performance. The experiment conducted in this study forms the basis of the findings, without any preconceived assumptions or performance expectations. The aim of this paper is to test the accuracy of the three LLMs in distinguishing factual based information from deceptive news by using a common prompt in a simulation. The evaluation is based on comparing their responses to a refined version of the established ISOT dataset [7]. The base ISOT dataset is a widely used benchmark dataset for fake news detection and consists of two main categories: Real News and Fake News. The creation of it was carried out by the ISOT research lab at the University of Victoria, specializing in Information Security and Object Technology. The dataset contains news articles that were collected from reputable news websites for the real news category and from unreliable sources for the fake news category. The dataset is primarily used for training and evaluating machine learning models designed to distinguish between fake and real news. The dataset includes thousands of labeled news articles, making it suitable for supervised learning tasks in natural language processing and machine learning research focused on fake news detection [8].

The main research questions are;

- RQ1: How accurately do the considered LLM’s classify true and fake news in a controlled simulation while dealing with large quantities of data?
- RQ2: How does their performance compare to natural language processing (NLP) models pretrained for fake news binary classification?

This study's primary aim is about the concept of "fake news," which is extensively explored within various interdisciplinary fields such as media studies, artificial intelligence, and education, further expanding the already widespread applications of LLMs [9].

Furthermore, this paper employs essential principles and concepts in its approach, incorporating techniques from psychology experimental designs and mathematical modeling. It encompasses both qualitative and quantitative viewpoints in analyzing and interpreting the data. In addition, this paper integrates essential principles and concepts into its methodology, including psychological experimental designs and mathematical models. This approach combines qualitative and quantitative viewpoints in the analysis and interpretation [3]. This research offers valuable insights into the potential use of LLMs as an effective solution for addressing the issue of cyber deceptions, including misinformation, disinformation, and the emergence of sophisticated fake news and AI-generated content. The paper highlights that relying solely on human skills and information literacy is inadequate in effectively addressing these falsehoods.

The paper is organized as follows: The following sections provide necessary summaries on various important topics to help readers understand the simulation process used in this study. These areas include (a) the basics of fake news testing and experimentation, (b) modeling cyber risks associated with misinformation and disinformation, and (c) a review of previous research on LLMs and their relation to fake news. The third chapter provides a detailed explanation of the simulation steps, allowing for replication of the experiment by interested readers. Additionally, this section addresses the limitations inherent in the simulation. The results of the simulation are then presented in a dedicated chapter, followed by an in-depth discussion of the findings. In conclusion, the paper provides recommendations and suggests potential areas for future research in this particular field.

2. RELATED WORK

2.1 Fake News Checking Methods

In the realm of misinformation and disinformation, vulnerability pertains to an individual's likelihood of being prone to accepting intentionally spread false news information. The main reason for conducting experiments to test agents' ability to combat deceptive social media content is the lack of publicly available comprehensive datasets that demonstrate this vulnerability [10]. Additionally, Facebook has taken actions to prohibit researchers engaged in similar activities, indicating that such endeavors violate the company's terms and policies. Critics assert that this decision was driven by the platform's desire to protect its public image and brand [11]. Consequently, future and previous experiments will continue to strive towards simulating a real social media platform through the inclusion of unmanufactured news articles, predominantly sourced from authentic social media content or reputable media organizations [10].

The actual tests consist of comparing true and fake pieces of news. Experimentation formats often differ, with controlled studies often conducting in-person tests. In these tests, participants are provided with printouts of news headlines and are asked to determine the authenticity of the information. In certain rigorous study designs, the use of electronic devices and internet access is strictly prohibited to authenticate the content of the test items. More contemporary variations of the study design have shifted towards electronic testing, commonly done online to reach a larger and more diverse group of participants [9].

2.2 Cyber Risk Modeling of Fake News

The performance simulations with agents generate two sets of performance data. The first set measures the accuracy of their performance, specifically the number of correct detections out of the total number of presented items, expressed as count data. The second set measures the time taken by the agents to complete their assessments, typically measured in seconds. These performance data are then analyzed to determine any possible correlations with a variety of potential predictors. The purpose of this analysis is to identify factors that may have influenced the observed performance gains or losses in the experimental data [9].

The areas that have been tested for their impact on susceptibility to mis/disinformation include demographic and socioeconomic factors such as age, gender [3, 12], family income, spoken language [12, 13], religion and military status [13, 14]. Other factors that have been examined include psychological perceptions and political positions, although these have not produced significant results so far [3, 15]. In the field of information sciences, the focus is often on the characteristics of the test items rather than the individuals themselves. Some of the predictors in this case could involve factors like the format or textual nature of news content, accompanying contextual hints, and metadata like the time it is circulated and the extent of its exposure to social media users [3]. When studying the behavioral sciences, factors that can help predict behavior include observing how individuals engage with social media content and identifying their preferred device type for browsing social media websites [13–15]. Some argue that the way in which the information environment is designed, including social media policies and terms of use agreements, can influence the dissemination of misleading or false information [16].

2.3 LLMs in Disinformation Mitigation

The proposal suggests that regular social media users, regardless of their level of information literacy, struggle to easily and rapidly identify specific types of misinformation. The philosophical basis of this idea is to use LLMs to address and counteract this issue [17]. These deceptive tactics have been specifically developed to pose significant cybersecurity risks, as they are meticulously designed to manipulate public opinions and foster divisions within society [10]. A significant assessment drew a comparison between deceptive tactics and other types of disruptive attacks [18]. These events often occur in large numbers and intensify during periods of significance, such as elections and public emergencies, which pose a threat to national security due to the potential for civil unrest [19].

Given the challenge of countering cyber deceptions fueled by advanced technologies, ongoing research suggests that employing similar technologies to combat these potent falsehoods is a recommended approach [7, 10]. Currently, the investigation into deepfakes, utilizing machine and deep learning, is a rapidly developing and fruitful area of research [20]. Moreover, the distribution of disinformation through bots poses a significant challenge for independent fact-checkers who are often ill-equipped to handle it alone. This is where LLMs play an important role in leveling the playing field within the information ecosystem [7].

2.4 Research Niche: Accuracy Comparison of Most Popular Large Language Models for Fake News Automatic Detection

Based on the previous facts provided, the following outcomes are apparent:

a. The *fake news verification model as a LLM based solution* proposed in this research may be used to combat similarly technologically powered deceptions has its merits and needs to be addressed.

b. Although existing studies with promising results or outcomes were initiated by different research teams that were taking interesting approaches like using sets of questions [21], involving AI generated fake content [22], against style-based attacks [19], or fine-tuning models like Llama 2 [23], or ChatGPT [24]. Such researches determined the ability of *LLM's as agents* for fact checking instead of persons but their designs and datasets or samples are still early and may be refined upon in future studies.

c. This article answers some of the limitations illustrated in previous research papers [7, 10, 20], and improves them by realizing a simulation involving a large amount of data [25], and increasing the number of models compared.

3. EXPERIMENTAL SETUP & SIMULATIONS

3.1 Overview

The approach used in this study is described in this chapter, along with the materials employed. The chapter covers various aspects such as the choice of LLM), the procedures for dataset gathering, the structure for simulations, the evaluation metrics, the process of analysis, and the experimental limitations in this study design. Moreover, the section discusses the considerations of reproducibility and ethics.

3.2 Selection of LLMs, NLPs and Short Descriptions

1) OpenAI's ChatGPT 4.0 and 4.0 beta: The 4.0 version was released in 2023 and introduces enhancements in both processing power and understanding. It offers better contextual awareness, reduces the frequency of errors, and provides more detailed, nuanced responses. Also, its more extensive contextual memory, allows it to manage longer conversations more effectively. The 4.0 Beta version includes experimental features, such as real-time web browsing capabilities and more dynamic input handling. It is designed to address some limitations of the standard 4.0 version by offering live access to web data, enabling it to provide more current and accurate information. The beta version also incorporates user feedback to refine its functionality and performance further [4].

The 4.0 version was added as a control variable, in order verify the improvements over the past models.

2) Meta's LLaMA 3.1 is an advanced LLM developed by Meta (formerly Facebook), released on the 23.07.2024, as part of its AI research efforts. Unlike ChatGPT, which is a proprietary model developed by OpenAI, LLaMA 3.1 is open source, which is particularly appealing to the academic and research communities. Furthermore, it is designed for efficiency in computational resources, allowing it to be fine-tuned with less computational power and should be able to deal with large amounts of data. One of the significant limitations of LLaMA 3.1, particularly in the European Union, is related to regulatory compliance. The EU's General Data Protection Regulation (GDPR) imposes rigorous demands on data privacy and processing, which poses challenges for deploying AI models like LLaMA 2. These regulations can restrict the use of data for training and limit the model's deployment in certain applications within the EU, necessitating additional compliance measures and potentially reducing the model's effectiveness compared to less-regulated environments. For this research the 3.1-8B-Instruct model was chosen [26]. The LLaMA 2 model was also used as a control measure for verifying the improvement of the 3.1 version over the 2 model.

3) Google's Gemini is developed by Google, designed to compete with other advanced AI models such as ChatGPT and LLaMA. Gemini is deeply integrated with Google's vast ecosystem, leveraging real-time data from Google's search engine, cloud services, and other platforms. This allows Gemini to provide more up-to-date and contextually relevant information compared to ChatGPT and LLaMA, which rely on static datasets with a fixed knowledge cutoff. The premium version of Gemini is also designed with multimodal capabilities, meaning it can understand and generate not only text but also images, audio, and potentially video. This makes the model more versatile in handling a broader range of tasks [27].

4) For further testing, a *Llama 3.1 8B model [26]*, was fine-tuned using the ISOT dataset, specifically for fake news fact-checking using an efficient methodology. The fine-tuning process employed Parameter-Efficient Fine-Tuning (PEFT) with Low-Rank Adaptation (Lo-RA) to selectively adjust the model's linear layers by adding additional parameters. This was carried out using mixed precision (FP16) to strike a balance between computational efficiency and accuracy, and Int8 quantization to reduce the model's size, making it suitable for deployment in environments with limited resources. The optimization process used the AdamW optimizer, with a linear learning rate scheduler starting at 0.00003, including a 10% warmup phase. Gradient accumulation steps were set to 4, with a maximum gradient norm of 1 to ensure stable training. The model was trained for 4 epochs using small batch sizes to enable gradual learning from the dataset, allowing it to effectively grasp the nuances of fake news detection. This approach ensured that the fine-tuned Llama 3 model was not only optimized for accuracy and performance but also ready for practical deployment in resource-constrained settings, capable of handling the complexities of fake news detection.

5) BERT (Bidirectional Encoder Representations from Transformers) [7], is the base version of Google‘s transformer-based model that understands context by looking at the surrounding words (both left and right) in a sentence. It is pre-trained on vast amounts of text data and can be fine-tuned for various NLP tasks, including fake news detection.

6) FakeBERT [28], is a specialized variant of BERT fine-tuned for the specific task of detecting fake news. By leveraging BERT's robust understanding of language and context, FakeBERT is optimized using the ISOT Dataset [25].

7) RoBERTa [29], is an advanced variant of the BERT (Bidirectional Encoder Representations from Transformers) model, RoBERTa (Robustly optimized BERT pretraining approach) improves on BERT by using a larger dataset and removing certain training constraints, such as the next sentence prediction task. It achieves state-of-the-art results in several NLP tasks.

8) ALBERT [30], or A Lite BERT is a scaled-down version of the original BERT model that reduces memory consumption and computational costs while maintaining high accuracy. ALBERT achieves this by using parameter sharing across layers and factorized embedding parameterization, making it more efficient for NLP tasks.

9) LSTM-RRN [31], or the Long Short-Term Memory model is a type of recurrent neural network (RNN) specifically designed to handle sequences of data. LSTM-RRN is a recurrent model that effectively retains long-term dependencies, making it well-suited for tasks like fake news detection by processing sequential text data and capturing contextual information.

10) BiLSTM-RRN [32], or Bidirectional LSTM-RRN is an extension of the standard LSTM, where two LSTMs are run in parallel, one processing the text from start to end and the other from end to start. This bidirectional approach enables the model to capture context from both past and future inputs, enhancing its performance in sequence-based tasks like fake news detection.

11) DeepFakE [33], is a deep learning model specifically designed for fake news detection, using advanced neural networks that capture deep contextual relationships within the text. It can differentiate between real and fake news by analyzing intricate patterns in the data, though its accuracy depends on the quality of the dataset, thus was selected for this test.

12) EchoFakeD [34], is another deep learning-based model designed to address misinformation and fake news by focusing on echo chambers and the way misinformation propagates in social media. It uses both textual analysis and user interaction patterns to detect and mitigate the spread of fake content.

3.3 Data Collection

3.3.1 The ISOT dataset

We used a fine-tuned version consisting of 34.098 pieces of news [25], based on a fact-checked and largely utilized dataset that has been mentioned in over 300 research between 2021 and 2022. The base dataset is called the ISOT Fake News Dataset [8], as provided by the site of the Department of Electrical & Computer Engineering within the University of Victoria, Canada [35]. The dataset has 2 sets of articles (17049 fake and 17049 real), gathered from Reuters and other media sources that have been labeled as true or false by PolitiFact or other fact-checking organizations.

3.3.2 Inclusion based on timeline

All the media texts used in the research are up to the year 2017 to level the playing field for all the LLM's tested and to be able to provide a baseline when comparing them to other fine-tuned natural language processing (NLP) models. In addition, the dataset we utilize exclude. The purpose of this is to avoid any potential prejudices in the simulation, which could be influenced by different forms of media organizations, including any type of profit entities (state, publicly or private held). In this study, we depend exclusively on independent outlets specialized in fact checking as the authority [7, 10].

3.3.3 Classification as true or false

The assigned label for the news within the dataset is binary (0) True, (1) False. The basis of this simulation relies solely on binary classification for increased accuracy, although human checked content usually fits into four categories True, Partially True, False, Partially False. Our research will exclude the partial categories because the labels ”Partially True” and ”Partially False” are unclear and could be confusing for LLMs. Our aim is to remove this ambiguity and clearly indicate that our dataset consists of a combination of true and false news. In practice, items labeled as ”Partially True” and ”Partially False” impose challenges for AI detection [10].

3.4 Experimental Setup

3.4.1 Details for the conditions of the setup

To guarantee a regulated assessment of the Language Learning Models (LLMs), a meticulous testing environment was deliberately created. Gemini and ChatGPT 4.0 beta were run their native environment while the LLaMA 3.1 models were run on hardware configurations provided through the huggingface.co platform. The same conditions were provided for the ChatGPT 4.0 and LLaMA 2 models to avoid any performance issues caused by hardware. To ensure a consistent environment for all models, a cloud-based platform was used, where all LLMs operated in identical conditions and utilized the same required libraries. By following this method, it ensured that any variations in LLM performance could be entirely ascribed to the models themselves without any external factors influencing it. The NLP models presented were trained for binary text classification (true or false) on the same ISOT dataset using the 80/20 ratio for training/testing.

3.4.2 Methods for preventing unaccounted variables

In order to avoid any potential impact of uncontrolled variables on the results, various control measures were adopted. One of these measures involved testing all the LLMs simultaneously, under identical conditions and during the same time of day, thereby mitigating any potential influence of network traffic or server load. Secondly, the same 2 files were presented to each LLM, either through cloud sharing (for the LLaMA 3.1 model), or through direct feeding of the files (ChatGPT 4.0 beta and Gemini), or by splitting the files into smaller parts (ChatGPT 4.0 and LLaMA 2) ensuring that all LLMs had to do the exact same evaluation. To preserve consistency, no updates or changes were allowed for the LLMs during the testing phase. The models underwent testing with prompts derived from two dataset files (true isot.csv and fake isot.csv). These prompts were created with the intention of generating a response that includes the percentage of either true or false news lines.

3.5 Metrics for the LLMs Evaluation

3.5.1 Accuracy (Acc)

This evaluation metric determines the accuracy of classifying instances as either true or fake news. It is computed by adding the number of true positives (Tp) and true negatives (Tn), then dividing it by the total number of instances (Total). This metric is valuable in cases where there is a balanced distribution of both true and fake news items. Acc = (Tp + Tn) / Total [7]. A second metric was not included due to the time limitations in this project, but it is a desiderate as a part of a future project. The chosen method of the analysis that each LLM will provide will be held for discussion in this paper as to clarify certain aspects.

3.5.2 Precision

Precision measures the proportion of true positives out of all predicted positive instances. In this context, it indicates how often the model correctly identifies fake news when it has classified an article as fake. Precision is important when the cost of false positives (incorrectly labeling real news as fake) is high. Precision= Tp/(Tp+Fp). A higher precision score indicates fewer false alarms and greater reliability when identifying fake news.

3.5.3 Recall

Also called Sensitivity or True Positive Rate, it measures the proportion of actual positive instances (fake news) that are correctly identified by the model. It provides insights into the model's ability to detect fake news among all fake news instances. This is relevant while the primary goal is to ensure that all fake news articles are flagged, even if occasionally misclassifying real news. Recall=Tp/(Tp+Fn). This metric is particularly useful in scenarios where missing fake news (false negatives) is highly detrimental.

3.5.4 F1-Score

This metric is the harmonic mean of precision and recall, providing a balanced measure when both metrics are equally important. It is useful when the cost of both false positives and false negatives needs to be minimized. F1-Score=2×(Precision×Recall)/(Precision+Recall). Such a metric helps to find the balance between precision and recall, offering a single metric that encapsulates both.

3.6 Limitations for the Research

1. To resolve any ambiguity in the classification of the LLMs, a requirement for the analysis was made to include two distinct choices (True / False), thus ensuring a more specific response.

2. Furthermore, our assessment criteria only focus on the precision of the LLMs' categorizations. Consequently, this research neglects other aspects of performance such as response time and the quality of the generated text.

3. We must acknowledge that the original fact-checking agencies are not immune to biases, errors and misclassifications. Incorrectly categorizing a news item could unfairly impact the evaluation of LLMs.

4. RESULTS AND ANALYSIS

4.1 LLMs Performance.

Response for (RQ1): Comparing the results of the 3 main Large Language Models (LLMs) in identifying fake news from true pieces of media from large datasets, OpenAI's GPT-4.0 beta and LLaMA 3.1 outperformed their predecessors GPT-4.0 and LLaMA 2 in being able to analyze large amounts of data. The average accuracy score for the tested LLMs was *55,2905%*, indicating a moderate ability to correctly identify true versus fake pieces of text. Gemini scored slightly higher than GPT-4.0 beta and Llama 3.1 and the fine-tuned version outperformed them overall suggesting a better analysis capacity and a capacity for improvement. ChatGPT had by far the highest score for the true news dataset and the worst score for the fake news dataset, suggesting a bias in identifying fake facts. It is worth mentioning that both LLaMA models had similar accuracy scores for both datasets, making it the most constant in this task. Although these models show they can effectively identify truthful information from deceptive news, the improvement shown in the fine-tuned version shows there is room for enhancement. The results show the significance of continuous research and improvement in extensive language models, particularly in the task of distinguishing facts from misinformation. Additionally, this study encourages further exploration into methods of better training these models to increase their precision in differentiating between accurate and deceitful data (TABLE 1).

Table 1: Individual Scores of the LLMs

LLM Name	True Positive/ Total Positive/	True Negative/ Total Negative	Total True/ Total	Accuracy	Precision	Recall	F1-Score
LLaMA 2	x/17049	x/17049	x/34098	NA	NA	NA	NA
LLaMA 3.1	8514/17049	8474/17049	16988/34098	49.8211%	0.498215	0.499384	0.498798
LLaMA 3.1 8B fine- tuned	10392/17049	9993/17049	20385/34098	59.7835%	0.595598	0.60953	0.602487
ChatGPT 4.0	x/17049	x/17049	x/34098	NA	NA	NA	NA
ChatGPT 4.0 beta	15728/17049	2803/17049	18531/34098	54.3462%	0.524721	0.922517	0.668949
Gemini	12034/17049	8046/17049	20080/34098	57.2112%	0.546081	0.705847	0.615770

4.2 Differences Between LLMs and Specialized NLP Models Trained on the ISOT Dataset for This Binary Task.

Response for (RQ2): The LLMs performance in accurately identifying true and fake news is somewhat successful but when compared with specialized Natural Language Processing (NLP) models specialized in text classification and trained on the same ISOT datasets, like Google's BERT variants or Long-Short-Term Memory Models the LLMs achievements are modest (TABLE 2).

Table 2: Individual performance of LLMs compared with eight deep learning/transformer-based NLP models in automatic fake news detection tasks on the ISOT dataset [7].

Model	Accuracy	Precision	Recall	F1-Score
RoBERTa	99,96%	0.99	0.99	0.99
LSTM-RRN	96,97%	0.97	0.97	0.97
BiLSTM-RRN	98,75%	0.97	0.97	0.97
ALBERT	97,80%	0.97	0.97	0.97
FakeBERT	98,74%	0.99	0.99	0.99
DeepFakE	88,64%	0.82	0.84	0.84
EchoFakeD	92,30%	0.90	0.86	0.88
BERT	98,13%	0.98	0.98	0.98
LLaMA 3.1	49,82%	0.49	0.49	0.49
LLaMA 3.1 8B fine- tuned	59,78%	0.59	0.60	0.60
ChatGPT 4.0 beta	54,34%	0.52	0.92	0.66
Gemini	57,21%	0.54	0.70	0.61

The comparison between AI and human-led fact-checking agencies underscores the strengths and weaknesses of AI in this field. It demonstrates the importance of continuous research and development to enhance AI's capacity to comprehend and analyze sophisticated information. Nevertheless, it also emphasizes that, currently, human fact-checkers remain more dependable in detecting deceitful news content.

5. CONCLUSION

This research emphasizes the potential benefits and limitations of Large Language Models (LLMs) in the ongoing battle against disinformation and misinformation. Our analysis of three prominent LLMs - OpenAI's ChatGPT 4.0 beta, Google's Gemini, and Meta's LLaMA 3.1 - revealed that they possess moderate capacity in distinguishing truth or falsehoods, with an average accuracy score of 55.29%. However, it is important to note that these LLM models still fall short in classification tasks compared to specialized Natural Language Processing models trained on the same dataset and the meticulous contextual analysis conducted by humans in reputable organizations like PolitiFact.

Nevertheless, we must not overlook the significance of the advancements made by these AI models. The progress demonstrated by the GPT and LLaMA models indicates a future where AI models will excel at accurately processing news and information. One reason for the LLMs underachievement lays in their employment of more rudimentary sentiment analysis tools (in the case of ChatGPT) or Transformer based Deep Learning-based Text Classification models (in the case of LLaMA 3.1).

The intersection of AI models and human expertise provides a profound consideration of the current era we live in. The utilization of AI as a potent weapon against misinformation signifies an important milestone, underscoring the indispensable importance of human cognition, discernment, and emotional intelligence. Consequently, rather than viewing the advancement of AI as a path towards rendering humans obsolete, it should be regarded as a chance for synergistic cooperation.

As we navigate an era defined by the importance of truthful information, our ability to combine technological advancements with our distinct human cognition becomes crucial for our well-being and progress. By synergistically integrating AI capabilities with our innate abilities, we can establish a formidable defense against the pervasive spread of misinformation, empowering truth to prevail over falsehoods and creating a future where honesty thrives.

References

[1] Quelle D, Bovet A. The Perils & Promises of Fact-Checking With Large Language Models. 2023. ArXiv Preprint: https://arxiv.org/pdf/2310.13549

[2] Kareem W, Abbas N. Fighting Lies With Intelligence: Using Large Language Models and Chain of Thoughts Technique to Combat Fake News. In: Bramer M, Stahl F, editors. Artificial Intelligence XL: 43rd SGAI International Conference on Artificial Intelligence. Springer. 2023;14381:253-258.

[3] Su J, Cardie C, Nakov P. Adapting Fake News Detection to the Era of Large Language Models. In: Findings of the association for computational linguistics: NAACL. Association for Computational Linguistics. 2024:1473-1490.

[4] Open AI. GPT-4 Technical Report. 2023. ArXiv preprint: https://arxiv.org/pdf/2303.08774

[5] Touvron H, Martin L, Stone K, Albert P, Almahairi A, et. al. Llama 2: Open and Efficient Foundation Language Models. 2023. ArXiv preprint: https://arxiv.org/pdf/2307.09288

[6] https://www.techradar.com/computing/artificial-intelligence/what-is-google-gemini

[7] Repede SE, Brad R. A comparison of artificial intelligence models used for fake news detection. Bulletin of “Carol I” National Defence University. 2023;12:114-131.

[8] Ahmed H, Traore I, Saad S. Detection of Online Fake News Using N-Gram Analysis and Machine Learning Techniques. International conference on intelligent, secure, and dependable systems in distributed and cloud environments. 2017:127-138.

[9] Caramancion KM. An Interdisciplinary Perspective on Mis/Disinformation Control. In: International Conference on Electrical Computer Communications and Mechatronics Engineering (ICECCME). IEEE PUBLICATIONS. 2023;2023:1-6.

[10] Repede SE. Researching Disinformation Using Artificial Intelligence Techniques: Challenges. Bulletin of “Carol I” National Defence University. 2023;12:69-85.

[11] https://www.npr.org/2021/08/04/1024791053/facebook-boots-nyu-disinformation-researchers-off-its-platform-and-critics-cry-f

[12] Blanco-Herrero D, Amores JJ, Sánchez-Holgado P. Citizen Perceptions of Fake News in Spain: Socioeconomic Demographic and Ideological Differences. Publications. 2021;9:35.

[13] Bryanov K, Vziatysheva V. Determinants of Individuals' Belief in Fake News: A Scoping Review Determinants of Belief in Fake News. PLOS ONE. 2021;16:e0253717.

[14] Maffioli EM, Gonzalez R. Are Socio-Demographic and Economic Characteristics Good Predictors of Misinformation During an Epidemic? PLOS Glob Public Health. 2022;2:e0000279.

[15] Valencia-Arias A, Arango-Botero DM, Cardona-Acevedo S, Paredes Delgado SS, Gallegos A. Are Socio-Demographic and Economic Characteristics Good Predictors of Misinformation During an Epidemics. Informatics. 2023;10:38.

[16] Caceres MF, Sosa JP, Lawrence JA, Sestacovschi C, Tidd-Johnson A, et. al. The Impact of Misinformation on the COVID-19 Pandemic. AIMS Public Health. 2022;9:262-277.

[17] Chen C, Shu K. Combating Misinformation in the Age of Llms: Opportunities and Challenges. AI Mag. 2024;45:354–368.

[18] https://www.cyber.gc.ca/en/guidance/national-cyber-threat-assessment-2023-2024

[19] Liu Y, Zhu J, Zhang K, Tang H, Zhang Y, et al. Detect Investigate Judge and Determine: A Novel LLM-Based Framework for Few-Shot Fake News Detection. 2024. ArXiv preprint :https://arxiv.org/pdf/2407.08952

[20] Repede SE. Some legislative and technological limitations in combating false information. Law and Life Conference Alba Iulia. Part of ISSN: 2587-4365. 2023:39-51.

[21] Caramancion KM. News Verifiers Showdown: A Comparative Performance Evaluation of ChatGPT 3.5, ChatGPT 4.0, Bing AI, and Bard in News Fact-Checking. 2023. ArXiv preprint: https://arxiv.org/pdf/2306.17176

[22] Wu J, Guo J, Hooi B. Fake News in Sheep's Clothing: Robust Fake News Detection Against LLM-Empowered Style Attacks. 2023. arXiv preprint: https://arxiv.org/pdf/2310.10830

[23] Pavlyshenko BM . Analysis of Disinformation and Fake News Detection Using Fine-Tuned Large Language Model. 2023. ArXiv preprint: https://arxiv.org/pdf/2309.04704

[24] Caramancion KM. Harnessing the Power of ChatGPT to Decimate Mis/Disinformation: Using ChatGPT for Fake News Detection. In: IEEE world AI IoT congress (AIIoT). IEEE Publications. 2023:42-46.

[25] https://huggingface.co/datasets/Phoenyx83/ISOT-Fake-News-Dataset-FineTuned-2022

[26] https://github.com/meta-llama/llama-models/blob/main/models/llama3_1/MODEL_CARD.md

[27] https://blog.google/technology/ai/google-gemini-ai/

[28] Kaliyar RK, Goswami A, Narang P. Fake Bert: Fake News Detection in Social Media With a Bert-Based Deep Learning Approach. Multimed Tools Appl. 2021;80:11765-11788.

[29] Liu Y, Ott M, Goyal N, Du J, Joshi M, et. al. Roberta: A Robustly Optimized Bert Pretraining Approach. 2019. Arxiv Preprint; https://arxiv.org/pdf/1907.11692

[30] Lan Z, Chen M, Goodman S, Gimpel K, Sharma P, et al. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. 2019.Arxiv Preprint: https://doi.org/10.48550/arXiv.1909.11942

[31] Wang WY. "Liar, Liar Pants on Fire": A New Benchmark Dataset for Fake News Detection. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, Vancouver, Canada. Association for Computational Linguistics. 2017;2:422–426.

[32] Bahad P, Saxena P, Kamal R. Fake News Detection Using Bi-directional Lstm-Recurrent Neural Network. Procedia Computer Science. 2019;165:74-82.

[33] Kanchana M, Kumar VM, Anish TP, Gopirajan P. Deep Fake BERT: Efficient Online Fake News Detection System, *International Conference on Networking and Communications, Chennai, India. (ICNWC)*. IEEE. 2023:1-6.

[34] Kaliyar RK, Goswami A, Narang P. Echo Fake D: Improving Fake News Detection in Social Media With an Efficient Deep Neural Network . Neural Comput Appl. 2021;33:8597-8613.

[35] https://drive.google.com/open?id=1IoTRrJNDJqvaG3hnUpnHQyGvPAJbO8y3