Fake News Detection With Advanced LLMs

Tyler Berg
University of Notre Dame
tberg3@nd.edu

Christian Northrup
University of Notre Dame
cnorthr2@nd.edu

Haomin Zhuang
University of Notre Dame
hzhuang2@nd.edu

Abstract

The rapid dissemination of fake news across digital platforms poses significant challenges to public trust, institutional stability, and informed decision-making. This study evaluates the performance of advanced Large Language Models (LLMs), specifically GPT-4o and LLaMA 3.2, in detecting fake news within a zero-shot framework. Using two benchmark datasets, FakeNewsNet and Snopes, we assess model performance through metrics such as accuracy, precision, recall, F1 score, and AUC-ROC. Our analysis reveals that GPT-4o demonstrates superior reliability and consistency compared to LLaMA 3.2, but both models exhibit notable limitations in recall, particularly with ambiguous and nuanced claims. The false negative reasoning analysis highlights recurring issues, such as rigid data interpretation, limited contextual understanding, and skepticism of uncommon scenarios. These findings underscore the need for hybrid approaches that integrate LLMs with external knowledge bases and domain-specific fine-tuning to improve misinformation detection systems. Future work should explore fine-tuning techniques, adversarial training, and expansion into multimedia and multilingual datasets to enhance robustness and applicability.

1 Problem Definition

The proliferation of fake news on digital platforms presents a significant societal challenge and is increasingly becoming a major global concern [7]. False or misleading information can spread rapidly, influencing public opinion, destabilizing institutions, and even affecting elections. Traditional methods of detecting fake news, such as rule-based systems or statistical models, struggle to keep up with the evolving nature of misinformation, which often leverages nuanced linguistic patterns, rapidly changing topics, and diverse cultural contexts.

Recent advancements in Natural Language Processing (NLP), particularly the development of Large Language Models (LLMs) such as GPT-4, BERT, and T5, introduce novel challenges in combating the generation and spread of misinformation, but they have also opened new avenues for addressing these issues [4]. LLMs are pre-trained on vast amounts of textual data, allowing them to capture complex semantic relationships and context within text. This makes them well-suited for the task of fake news detection, where subtle distinctions between truthful and deceptive content are crucial.

The problem we aim to address in this study is how to effectively harness LLMs to detect fake news with high accuracy, while minimizing false positives and false negatives. Specifically, we define the problem as a supervised learning task where the input is a news article or social media post, and the output is a binary classification: either fake or real. This classification task involves the following challenges:

Fact-checking Capabilities: Detecting fake news requires models to be able to fact-check text. They must either verify if a statement is true or otherwise determine that it is false. It can also often require knowledge of information from recent events.

Limited Labeled Data: Although LLMs are trained on large corpora, labeled datasets for fake news detection are often limited in size and scope, making it challenging to fine-tune models for this task.

Bias and Misinformation: LLMs themselves may learn biases present in the data they are trained on, potentially affecting their ability to differentiate between credible and misleading information.

We aim to assess current state-of-art models such as GPT-4o and LLaMA 3.2 on their ability to detect misinformation. We also want to explore the impact of data quality and diversity on detection accuracy, ensuring that the model remains robust across different contexts and platforms.

2 Motivation

The spread of fake news has become a pressing issue in today's digitally connected world. With the rapid dissemination of information across social media platforms, it is increasingly difficult for individuals to distinguish between credible news and fabricated content. The ability of some language models to quickly generate disinformation lowers the barrier to entry for creating and spreading fake news. This can have serious societal implications, from eroding public trust in institutions to influencing political outcomes. Given these high stakes, developing robust methods for detecting fake news is a critical task for safeguarding the integrity of information.

The consequences of unchecked misinformation are severe. Fake news can mislead large audiences, exacerbate political divisions, and create confusion during crises. Traditional fact-checking mechanisms, which often rely on human oversight, cannot scale to meet the volume and speed at which misinformation spreads. Therefore, automatic detection systems powered by AI, specifically Large Language Models (LLMs), offer a promising solution to combat fake news in real-time, at scale. By leveraging these models' ability to understand and generate language, we can better identify misleading narratives and mitigate their impact.

While the use of LLMs in detecting fake news is gaining traction, there is still a lack of comprehensive evaluation across different model architectures and versions. For instance, new models like LLaMA 3.2 have shown advancements in language understanding and generation but have not been thoroughly assessed in the context of misinformation detection. Evaluating such models allows us to explore their strengths and weaknesses in handling this specific task, offering insights into whether newer architectures bring tangible improvements over earlier models like GPT-4 or BERT in detecting nuanced fake news content.

Existing research on fake news detection has made significant progress but is often limited in scope. Many prior studies have focused on small, domain-specific datasets or have not fully addressed the challenges of generalization across different types of misinformation. In contrast, our work aims to extend the scope of fake news detection by using more comprehensive and diverse datasets, ensuring that the model can generalize across a variety of topics and platforms. Through our evaluation, we aim to uncover key insights about the models' behavior in detecting fake news.

3 Related Works

The detection of misinformation has been an active area of research, with growing interest in leveraging large language models (LLMs) to automate and scale fact-checking efforts. Several recent works have explored the potential of LLMs to detect fake news and augment human fact-checkers.

Hu et al. (2024) [3] investigated the role of large language models, specifically GPT-3.5, in fake news detection. The study found that while GPT-3.5 could provide useful multi-perspective rationales, it underperformed when compared to smaller, fine-tuned models like BERT. The authors proposed a hybrid approach, where LLMs serve as advisors to smaller models by generating informative rationales, but they emphasized that LLMs alone are not yet reliable enough for direct misinformation detection. This study highlighted the limitations of LLMs in current applications but also pointed to their potential for improving fake news detection when integrated into hybrid frameworks.

Choi and Ferrara (2024) [2] introduced FACT-GPT, a framework aimed at improving the claim-matching phase of fact-checking using LLMs. Their approach leverages GPT models to match new claims to previously fact-checked ones, automating a key part of the fact-checking process. Their evaluation demonstrated that fine-tuned LLMs could match human-level performance in identifying false claims across large datasets, providing an efficient tool for claim matching. However, the paper primarily focused on fine-tuning LLMs for task-specific applications and did not explore their zero-shot capabilities on other datasets.

Boissonneault and Hensen (2024) [1] conducted an empirical evaluation of ChatGPT and Google Gemini models on the LIAR dataset for fake news detection. Their results showed high performance metrics such as accuracy, precision, and F1 score, but their study was limited to the LIAR dataset and did not evaluate model performance on other misinformation datasets. The authors also emphasized the importance of ongoing research into the strengths and weaknesses of LLMs to improve their robustness in real-world applications.

Wu et al. (2024) [6] found that fake news could be easily transformed in style by LLMs in order to evade state-of-the-art detectors. In order to make models more robust to such adversarial attacks, the authors developed SheepDog, a fake news detector that focuses more on the content of text than the style. SheepDog was superior in performance to both GPT3.5 and Llama 2 models. As it was tested on both Politifact and GossipCop, we could compare how the models we test compare to the performance of SheepDog.

Building on this prior work, our project aims to expand the evaluation of LLMs in fake news detection by focusing on GPT-4o and LLaMA 3.2. Unlike previous studies that often fine-tune LLMs for specific tasks, we will evaluate these models purely in a zero-shot setting.

4 Research Objectives

To address the challenges posed by misinformation, our study aims to:

Evaluate LLMs in a Zero-Shot Setting

Zero-shot learning allows models to perform tasks without specific training for that task. We test GPT-4o and LLaMA 3.2 on FakeNewsNet and Snopes to evaluate their ability to generalize.

Measure Detection Performance

We assess the performance of the models using standard evaluation metrics: accuracy, precision, recall, F1 score, and AUC-ROC. These metrics provide a comprehensive understanding of the models' strengths and weaknesses.

Understand Model Limitations

The project investigates limitations, such as hallucinations, bias, and adversarial susceptibility. Identifying these issues is critical for improving future models.

Compare Architectures

By comparing GPT-4o and LLaMA 3.2, we highlight the strengths and weaknesses of different architectures, providing insights into their suitability for misinformation detection.

5 Methodology

5.1 Datasets

Two datasets were used to evaluate the models:

- **FakeNewsNet:** This dataset includes news articles collected from the fact-checking website Politifact and that are enriched with social and temporal context, enabling nuanced evaluation.
- **Snopes:** A comprehensive repository of fact-checked claims, providing a benchmark for evaluating misinformation detection capabilities.

5.2 Models and Setup

GPT-4o: Known for its advanced semantic understanding, GPT-4o represents a robust choice for misinformation detection.

LLaMA 3.2: With improvements in efficiency and contextual understanding, LLaMA 3.2 provides a modern alternative.

Zero-Shot Framework: We employ zero-shot evaluation, testing models without fine-tuning to assess their generalizability across datasets.

6 Evaluation Method

To assess the performance of our fake news detection model, we employ a variety of standard evaluation metrics that provide a comprehensive understanding of the model's ability to correctly classify news articles as real or fake. The primary metrics used in this study include accuracy, precision, recall, F1 score, and the Area Under the Receiver Operating Characteristic Curve (AUC-ROC). Each metric offers different insights into the performance of the model, particularly in handling the balance between false positives and false negatives, which is critical in tasks like fake news detection.

Accuracy. Accuracy is a straightforward measure of how often the model's predictions are correct. It is calculated as the ratio of correct predictions (both true positives and true negatives) to the total number of predictions. Although accuracy is a useful baseline metric, it may not fully reflect the model's performance when dealing with imbalanced datasets, which is often the case in fake news detection where there may be more real news samples than fake news samples.

$$\text{Accuracy} = \frac{\text{True Positives} + \text{True Negatives}}{\text{Total Samples}} \tag{1}$$

Precision. Precision measures the proportion of correctly predicted positive instances (fake news) out of all instances that were predicted as positive. It highlights the model's ability to avoid false positives, which in the context of fake news detection, corresponds to wrongly classifying real news as fake. High precision is important to ensure that credible news is not incorrectly flagged as misinformation.

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \tag{2}$$

Recall (Sensitivity). Recall, or sensitivity, focuses on the model's ability to correctly identify positive instances. It measures the proportion of actual fake news samples that were correctly identified by the model. A high recall score indicates that the model is effectively catching most instances of fake news, although it may come at the expense of increasing false positives.

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \tag{3}$$

F1 Score. The F1 score is the harmonic mean of precision and recall, offering a balance between the two. It is particularly useful when dealing with imbalanced datasets, as it provides a single metric that accounts for both false positives and false negatives. A higher F1 score indicates that the model is effectively managing the trade-off between precision and recall, making it a more reliable metric in fake news detection where both types of errors (false positives and false negatives) carry significant consequences.

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

AUC-ROC (Area Under the ROC Curve). The AUC-ROC curve is a graphical representation of the trade-off between the true positive rate (recall) and the false positive rate across various classification thresholds. The Area Under the ROC Curve (AUC) provides a single score that summarizes the model's ability to distinguish between real and fake news across all thresholds. A higher AUC score indicates that the model is better at distinguishing between the two classes, with a value of 1 representing perfect classification and 0.5 indicating random guessing. AUC-ROC is especially valuable in cases where the class distribution is imbalanced, as it evaluates the model's performance across a range of decision boundaries.

$$\text{AUC} = \int_0^1 \text{ROC Curve}(t)\, dt \tag{5}$$

7 Experiments

7.1 Settings

Model. To evaluate the performance of leading models, we include ChatGPT and Llama in our experiments. ChatGPT, representing the most powerful closed-source model, is tested using GPT-4o, a cost-effective variant. For the open-source Llama series, we use Llama 3.2 11B, the latest version recently released. All experiments are conducted using APIs provided by OpenAI and Meta. For all hyperparameters, we keep them as the default within API settings.

Dataset. We evaluate the models using two datasets: FakeNewsNet and Snopes, as mentioned earlier. Each dataset offers unique characteristics and contributes to a comprehensive evaluation of the models' performance in detecting fake news. FakeNewsNet dataset is widely regarded as a baseline in fake news detection research. It comprises news articles and their associated labels, categorizing information as either true or false. FakeNewsNet is constructed from a variety of sources and has been used extensively in prior studies, making it a standard benchmark for comparing model performance in this domain. Its diverse and well-annotated content provides a solid foundation for evaluating fake news detection capabilities. Snopes dataset is a more recently published resource, offering the latest and most up-to-date information on misinformation and fact-checked content. Built from the Snopes fact-checking platform, it contains curated entries on a wide range of topics, from viral rumors to policy-related claims. The labels in Snopes are more broad but we fine-tuned it down into just true and false for our study leaving us with over 13,000 claims to test. The dataset's focus on current and emerging topics ensures relevance in real-world applications. In our experimental setup, we exclusively perform zero-shot evaluations on both datasets. Zero-shot evaluation aligns closely with real-world user scenarios, where models are required to generalize to unseen data without prior task-specific fine-tuning. This approach not only tests the intrinsic generalization capabilities of the models but also simulates practical deployment conditions, emphasizing their ability to provide reliable results in dynamic and unpredictable environments.

Model-Dataset Combination	Accuracy	Precision	Recall	F1 Score	AUC-ROC
GPT-4o on Snopes	0.80	0.78	0.82	0.80	0.85
GPT-4o on FakeNewsNet	0.70	0.72	0.68	0.70	0.64
LLaMA 3.2 on Snopes	0.65	0.62	0.68	0.65	0.60
LLaMA 3.2 on FakeNewsNet	0.60	0.58	0.62	0.60	0.55

Table 1: Performance Metrics for GPT-4o and LLaMA 3.2

8 Results and Analysis

Discussion. The visual results (Figures 1–4) highlight key differences in model performance across datasets and metrics. GPT-4o outperformed LLaMA 3.2 across all metrics, demonstrating greater reliability and consistency. However, both models showed significant variability between datasets, with higher performance on Snopes compared to FakeNewsNet. This discrepancy suggests that dataset structure and quality strongly influence model effectiveness.

A closer inspection of false negatives revealed key patterns. For example, ambiguous claims, such as those lacking proper context, were often misclassified. Additionally, GPT-4o demonstrated occasional over-reliance on linguistic patterns without verifying factual content. This aligns with the explanations retrieved from the code, where the models struggled with nuanced statements.

The ROC curve analysis (Figures 5–8), which underpins the AUC-ROC values shown in the metrics table, demonstrated that GPT-4o consistently achieved better discrimination capabilities compared to LLaMA 3.2, as evidenced by its higher AUC-ROC values. The false negatives collected in the code pipeline (referenced in Figure 4) often highlighted gaps in factual knowledge integration, suggesting that the models would benefit from hybrid approaches combining LLMs with external databases.

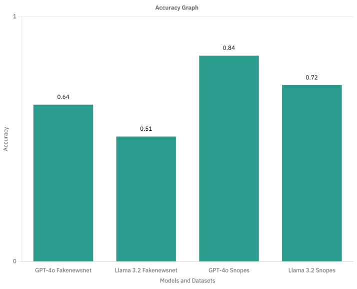

Figure 1: Accuracy Comparison Across Models and Datasets

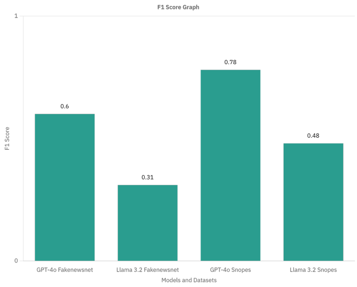

Figure 2: F1 Score Comparison Across Models and Datasets

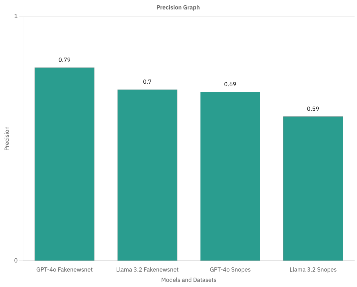

Figure 3: Precision Comparison Across Models and Datasets

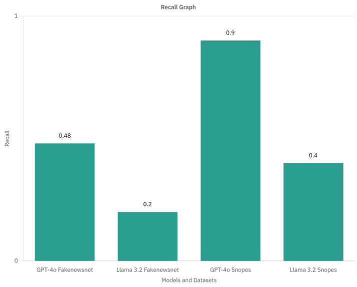

Figure 4: Recall Comparison Across Models and Datasets

On the FakeNewsNet dataset, both GPT-4o and LLaMA 3.2 underperform state-of-the-art machine learning models. Such models have achieved as high as 0.82 accuracy [5]. However, those machine learning models were specifically trained on and for the FakeNewsNet dataset, whereas GPT-4o and LLaMA 3.2 are general-purposed LLMs designed to perform across a wide range of tasks. While LLMs may have a performance tradeoff, they may be more generalizable to a wider variety of misinformation detection tasks. They also have the advantage of being interpretable, as LLMs can output reasons and explanations for its classifications.

8.1 False Negative Reasoning Analysis.

The recall performance of both GPT-4o and LLaMA 3.2 was notably suboptimal, as shown in Figure 4, highlighting a key area of concern. In tasks such as misinformation detection, low recall means that a significant proportion of actual fake news or valid claims are misclassified, which could have serious implications. Misclassifications could propagate false narratives or fail to highlight true claims, undermining the credibility and utility of such systems. Understanding why these errors occur is critical for enhancing the reliability of these models.

LLaMA's classification errors on claims that were actually true revealed several key patterns. One prominent issue was its over-reliance on literal interpretations and exact statistical data. For example, in evaluating the ratio of COVID-19 deaths to the U.S. population, LLaMA deemed a claim false because the calculated ratio did not perfectly match the claim, even though the approximation was reasonable in context. Similarly, it misinterpreted claims involving percentages or population estimates, such as the Ohio River Basin population, by rigidly applying its understanding of demographic data without accounting for alternative interpretations.

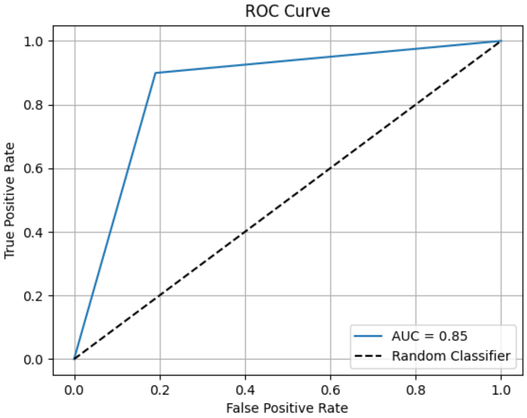

Figure 5: AUC-ROC Curve GPT-4o Snopes Dataset

Another significant source of error stemmed from LLaMA's challenges in handling ambiguous or incomplete evidence. For instance, in the claim about Grover Cleveland and the $1,000 bill, LLaMA failed to reconcile the historical context of the bill's issuance with its literal interpretation, incorrectly deeming the claim false. Similarly, in a claim about early New York Times reports on Adolf Hitler, LLaMA dismissed the claim based on

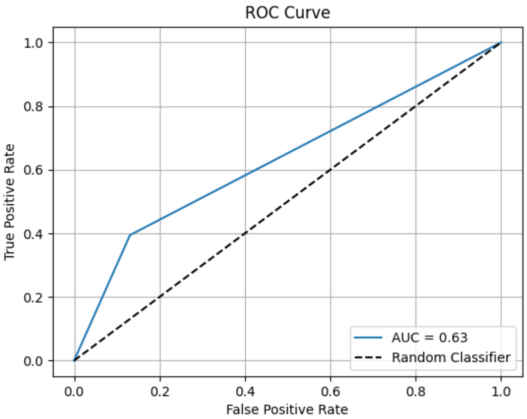

Figure 6: AUC-ROC Curve LLaMA Snopes Dataset

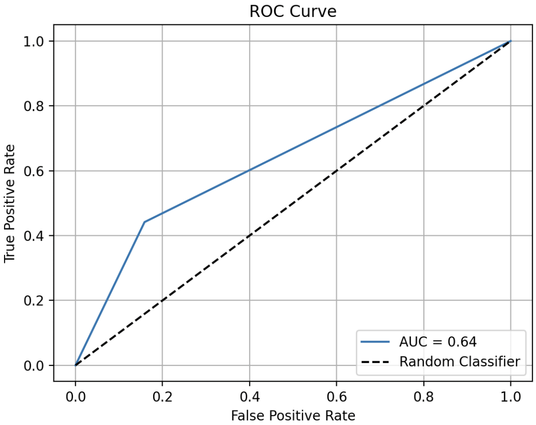

Figure 7: AUC-ROC Curve GPT-4o Fakenewsnet Dataset

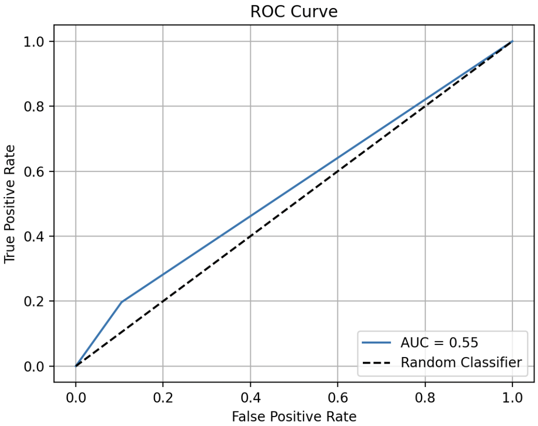

Figure 8: AUC-ROC Curve LLaMa Fakenewsnet Dataset

incomplete or contradictory evidence in its training corpus.

Contextual nuances also posed challenges. For example, LLaMA misclassified a claim about a viral video due to a narrow interpretation of the term "viral," focusing on modern definitions while ignoring the historical significance of the content. In another case involving the weight of a 14-pound opossum, the model relied heavily on general probabilities rather than considering the plausibility of an exceptional scenario.

Finally, LLaMA exhibited over-skepticism toward less common or extraordinary events. This was evident in claims about unusual phenomena or phrasing, where the model defaulted to false classifications without robust evidence. For instance, in a claim regarding coal burning's historical contributions to global warming, LLaMA dismissed the claim due to gaps in reconciling historical timelines with modern scientific interpretations.

GPT-4o, while performing better overall, also displayed notable reasoning flaws in its false negative classifications. For instance, it misclassified the statement "Earthquakes can influence volcanic activity" as false, explaining that there is no definitive evidence to confirm this relationship, despite ample scientific discussion suggesting plausible connections. Similarly, GPT-4o dismissed the claim "The tallest mountain in North America is Denali" by incorrectly prioritizing alternative naming conventions and geographic debates over widely accepted facts.

In another case, GPT-4o incorrectly classified the statement "Certain whales can live over 200 years" as false, arguing that such extreme lifespans are not biologically plausible, even though scientific evidence supports the claim for bowhead whales. Furthermore, GPT-4o dismissed the statement "The first email was sent in 1971" due to a lack of corroborating historical data in its training corpus, overlooking widely documented accounts of early email development.

These examples illustrate that while GPT-4o relies on a deeper contextual understanding compared to LLaMA, it can still falter in cases requiring flexible reasoning or evidence integration. Both models demonstrate tendencies toward rigid data interpretation, limited contextual understanding, and skepticism of uncommon claims. Addressing these shortcomings is essential for improving the efficacy of LLMs in misinformation detection.

9 Case Study

An advantage of using LLMs to detect fake news is that they can be prompted to explain the reasoning behind classifications. In this section, we focus on how can LLMs classify the news and dive into the failures. Understanding the reasons for both successful and unsuccessful classifications provides insight into the strengths and weaknesses of a model. We choose ChatGPT-4o here as it outperforms Llama in both datasets.

Case Study

News: Ten percent of the U.S. population lives in the Ohio River Basin, within which the East Palestine train derailment and chemical release occurred.

Model Answer: Fake news

Ground Truth: Real news

Model Explanation: The statement is false because the Ohio River Basin does not house 10% of the U.S. population. The basin covers parts of 15 U.S. states, but the total population in this area is less than 10% of the U.S. total population.

The example above demonstrates that LLMs evaluate news primarily based on their pre-trained knowledge. However, the explanations they provide often lack detailed reasoning steps or concrete evidence to support their conclusions. This observation highlights a critical limitation: LLMs tend to classify news using their pre-existing knowledge and general understanding rather than engaging in systematic reasoning or relying on verifiable evidence. As a result, they fall short of being reliable tools for detecting fake news. Moreover, this approach makes LLMs vulnerable to errors, particularly when confronted with the latest developments or events that occurred after their training data cutoff. Additionally, they may struggle to identify absurd or intentionally misleading fake news propagated by malicious sources, further undermining their effectiveness in such contexts.

Case Study

News: Anne Frank said: "Look at how a single candle can both defy and define the darkness."

Model Answer: Real news

Ground Truth: Fake news

Model Explanation: This quote is attributed to Anne Frank and reflects her profound insight into the power of hope and resilience.

The above example shows another weakness of using LLMs to detect fake news. Within its training data, the quote has been mistakenly attributed so often that the model just assumes that the statement is true. A more rigorous fact-checking model would attempt to actually verify the claim, rather than relying on a training dataset that may contain errors.

Case Study

News: Video clip captures NFL kicker Shaun Suisham booting a 110-yard practice field goal.

Model Answer: Fake news

Ground Truth: Fake news

Model Explanation: An NFL field is 100 yards long between the goal lines, with each end zone adding 10 yards. Therefore, a 110-yard field goal would be physically impossible even in a practice context, as it would require kicking the ball from well behind the field of play. The longest successful field goal in NFL history is 66 yards, achieved by Justin Tucker in 2021.

In this example, GPT-4o uses both logic and a known counterexample to successfully identify fake news. It correctly determines that kicking a field goal 110 yards is physically impossible. It also knows the distance of longest field goal, which additionally serves as evidence against the claim.

10 Conclusions

This study demonstrates the potential of LLMs for zero-shot misinformation detection and highlights both their strengths and limitations. GPT-4o emerged as the more reliable model, outperforming LLaMA 3.2 across all evaluated metrics, including accuracy, precision, recall, F1 score, and AUC-ROC. This corroborates recent research that has found a strong positive correlation between model size and fake news detection accuracy [4]. However, the analysis revealed that both models struggled with recall, indicating a significant challenge in identifying all instances of fake news or valid claims. This limitation has critical implications, as low recall can lead to the propagation of false narratives or the failure to flag accurate claims, undermining trust in these systems.

Error analysis provided further insights into these performance gaps. LLaMA often displayed a rigid reliance on literal interpretations and statistical precision, leading to misclassifications of claims involving approximate data or nuanced contexts. For example, claims related to population statistics, historical events, or extraordinary phenomena were frequently dismissed due to overly narrow interpretations. Similarly, GPT-4o, while leveraging a deeper contextual understanding, faltered on claims requiring flexible reasoning or integration of external evidence. It exhibited skepticism toward uncommon scenarios and occasionally failed to reconcile conflicting evidence within its training corpus. Examples include misclassifying claims about historical innovations, biological lifespans, and early communication technologies.

The study underscores the importance of addressing these shortcomings to improve the efficacy of LLMs in detecting misinformation. Hybrid approaches that combine LLMs with external knowledge bases or rule-based systems could enhance fact-checking capabilities, bridging the gap between linguistic reasoning and factual validation. Moreover, adversarial training techniques may improve robustness against deceptive inputs, while expanding datasets to include multimedia content and multilingual misinformation would test the models' ability to generalize across diverse formats and cultural contexts.

In conclusion, while LLMs like GPT-4o and LLaMA 3.2 show promise for scalable fake news detection, their limitations in recall, contextual reasoning, and error handling highlight the need for further innovation. Future research should focus on fine-tuning models, integrating hybrid architectures, and exploring adversarial defenses to create more reliable and adaptable misinformation detection systems.

11 Future Work

Future research must address several key areas to enhance the effectiveness of fake news detection systems. Fine-tuning experiments on challenging datasets could reveal pathways for improving model performance in domain-specific contexts. Such efforts would allow models to better understand nuanced claims and adapt to various misinformation scenarios. Additionally, hybrid models that integrate LLMs with external knowledge bases or rule-based systems may enhance fact-checking capabilities by bridging the gap between linguistic reasoning and factual validation. This could especially help to ensure that detection methods are up-to-date with the latest facts.

Expanding the scope of datasets to include multimedia content and non-English misinformation is another crucial avenue. This expansion would test the models' ability to generalize across diverse information formats and cultural contexts, ensuring a broader applicability of the detection system. Finally, adversarial training techniques could improve model robustness against intentionally deceptive or adversarial inputs, a growing challenge in misinformation spread. Together, these advancements would contribute to more accurate, reliable, and scalable fake news detection solutions.

References

[1] David Boissonneault and Emily Hensen. Fake news detection with large language models on the liar dataset, 2024.

[2] Eun Cheol Choi and Emilio Ferrara. Automated claim matching with large language models: Empowering fact-checkers in the fight against misinformation. In *Companion Proceedings of the ACM Web Conference 2024*, WWW '24, page 1441–1449, New York, NY, USA, 2024. Association for Computing Machinery.

[3] Beizhe Hu, Qiang Sheng, Juan Cao, Yuhui Shi, Yang Li, Danding Wang, and Peng Qi. Bad actor, good advisor: Exploring the role of large language models in fake news detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(20):22105–22113, March 2024.

[4] Dorsaf Sallami, Yuan-Chen Chang, and Esma Aimeur. From deception to detection: The dual roles of large language models in fake news. *10.48550/arXiv.2409.17416*, 2024.

[5] Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. Fakenewsnet: A data repository with news content, social context and spatialtemporal information for studying fake news on social media, 2019.

[6] Jiaying Wu, Jiafeng Guo, and Bryan Hooi. Fake news in sheep's clothing: Robust fake news detection against llm-empowered style attacks. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 3367–3378, 2024.

[7] Nurdiana Fariha 'Ainuddin, Nur Aimi Atirah Abdul Malik, Muhammad Izmeer Adil Aruan, and Salliza Md Radzi. Fake news and disinformation: Ethical impacts and responsibilities. *Journal of Islamic, Social, Economics and Development (JISED)*, 2023.