PERFORMANCE COMPARISON OF ZERO-SHOT AND TWO-SHOT PROMPTING IN DETECTING FAKE NEWS USING LARGE LANGUAGE MODELS

Muhammad Naim Syahmi Roslan[1], Masnizah Mohd[2*]

[1, 2] Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia,
43600 Bangi, Selangor, Malaysia

Emails: p127137@siswa.ukm.edu.my[1], masnizah.mohd@ukm.edu.my[2*]

ABSTRACT

Fake news detection is a highly crucial challenge in Natural Language Processing (NLP), particularly during significant social events like elections and national crises. This study uses the GPT-3.5-Turbo model to test the effectiveness of zero-shot and two-shot prompting in detecting fake news on the PolitiFact and Liar datasets. Zero-shot prompting consists of task instructions without examples, whereas two-shot prompting contains a few task-related examples. The methodology includes dataset preparation, Large Language Models (LLMs) response collection, encoding, and evaluation using metrics such as accuracy, precision, recall, and F1-score. The results show that two-shot prompting increases performance marginally across all parameters when compared to zero-shot prompting. PolitiFact's accuracy improved from 0.286 to 0.293, while Liar's improved from 0.220 to 0.226. Precision, recall, and F1-score also showed minor gains. However, these advances were not statistically significant and highlight the model's difficulty with handling multi-class classification in the political domain. The GPT-3.5-Turbo model performed better on the PolitiFact dataset, suggesting variability in performance across different datasets. In conclusion, although two-shot prompting provides a slight advantage, the GPT-3.5-Turbo's overall performance remains limited, indicating the need for more sophisticated techniques (such as advanced prompting methods or more powerful LLMs) to enhance fake news detection.

Keywords: Fake news detection; Natural language processing; Zero-shot prompting; Few-shot prompting.

1.0 INTRODUCTION

Fake news detection (FND) is a popular domain in NLP and is highly crucial as it can have significant impacts on society, especially during important events such as elections or national crises. For instance, misinformation in fake news can sway public opinion during elections and fuel distrust of vaccines, as seen during the recent global COVID-19 pandemic [1], [2]. Due to the widespread adoption of social platforms for information exchange, even tools originally designed for academic or workplace collaboration such as blogs or Facebook can shape knowledge construction and may inadvertently facilitate the spread of misinformation [3].

Traditionally, automated FND research has relied on supervised Machine Learning (ML) or Deep Learning (DL) algorithms. Despite the modernity of DL approaches, ML methods are still prevalent in FND. Traditional ML models like Logistic Regression (LR), Random Forest (RF), and XGBoost (XGB) have been evaluated alongside DL models such as CNN, BILSTM, and BERT [4]. Studies have shown that traditional ML models like Decision Tree, RF and LR can deliver high accuracy, with Decision Tree reaching 99.38% accuracy in one study [5]. In comparison, another study [6] found that LSTM models excelled with an average accuracy of 94.21%. The effectiveness of different algorithms varies based on the dataset and features used. While some research [5] found that simpler algorithms performed better, others [4], [6] highlight the benefits of advanced DL techniques. These findings suggest that traditional ML algorithms still hold significant value and produce results comparable to the newer and more powerful DL algorithms in FND tasks.

In recent years, high demand for AI research has resulted in the development of large language models (LLMs). These LLMs are trained on significant amounts of web data and are able to generate human-like and high-quality text. The most popular example of commercial LLMs is ChatGPT which utilizes LLMs from the GPT-series [7], [8], [9]. ChatGPT demonstrates the ability to generate text pertaining to any domain queried by the user. However, the veracity of the generated text is often inconsistent. Additionally, these recent advancements in LLMs have raised concerns about their potential to generate fake news content that can be indistinguishable from real news [10].

Several interrelated concepts regarding LLMs are in-context learning (ICL), zero-shot prompting, and few-shot prompting. ICL involves the ability of LLMs to learn and perform new tasks based on a few examples or instructions provided in the input prompt, without the need for additional training or fine-tuning [11], [12]. ICL is in contrast to standard machine learning (ML) methods, which involve significant training data and fine-tuning for every new task. Zero-shot prompting is giving the LLM only a written explanation of the task, with no examples [13]. The LLM is then expected to comprehend the problem and produce relevant results based on its general knowledge and abilities. In contrast, few-shot prompting involves giving the LLM a number of examples of the desired task along with the task instructions. The LLM can then use these few samples to better grasp the problem and produce more accurate results [14]. Research has also suggested that by increasing the number of examples (shots), it could lead to an increased Theory-of-Mind (ToM) performance in LLMs [11].

Even though LLMs could be used to generate fake news, its vast capabilities could also potentially be used to detecting them. LLMs have demonstrated a capability for generating factual information. As highlighted by [15], robust credibility assessment is essential in today's digital environment. The factual information generated by LLMs can be used to assess the credibility of news content. Real-world applications include using LLMs to scan articles or social media posts, extract key claims, and verify them against trusted sources such as established news outlets or fact-checking services. Another example is chatbots powered by LLMs, which can help clarify news content by explaining questionable claims, highlighting opposing evidence, and suggesting reliable sources for further verification. Researchers have explored whether this capability can be leveraged for FND. Research indicates that while LLMs can provide a multi-perspective approach to FND and expose falsehoods, they may still underperform compared to fine-tuned smaller language models (SLMs) in terms of carefully deliberating all the information required to reach a conclusion [16]. Though, this underperformance did not hamper interest in using LLMs for FND research. One popular approach is prompt-based tuning, which involves designing prompts in a specific way to elicit domain-specific knowledge from pre-trained LLMs. For instance, a context-rich prompt-based template can be used to extract targeted knowledge from pre-trained LLMs, thereby improving early detection of fake news [14]. This method has also shown significant improvements in few-shot learning settings, where limited data is available [14]. Additionally, few-shot learning techniques, like "Prompt-and-Align", utilize pre-trained LLMs and the structure of social contexts to identify fake news, even with a limited amount of labelled data [17]. This few-shot approach addresses the issue of label scarcity by embedding news articles within task-specific textual prompts, allowing the LLM to generate relevant knowledge for the task [17]. Despite the more popular usage of few-shot methods, zero-shot methods are also explored. For example, zero-shot methods such as "DetectGPT" [18] utilize the probability function structure of an LLM to identify machine-generated text such as fake news articles.

As examined in previous works, zero-shot and few-shot styles of prompting have emerged as promising techniques for detecting fake news using LLMs [14], [19]. Because it needs no additional training, the zero-shot method can adapt quickly to evolving fake news narratives, making it particularly useful in rapidly changing contexts. In contrast, the few-shot approach, such as two-shot prompting (where two task-related examples are provided in the input prompt) can help the LLM better understand the task and improve its performance, especially in cases where the fake news content follows a more discernible pattern that the zero-shot approach may struggle with.

Comparing zero-shot and few-shot prompting is crucial for helping researchers and practitioners understand the trade-offs and select the most suitable approach for their specific use cases. Additionally, it can provide valuable insights on the robustness and generalization capabilities of LLMs when faced with insufficient training data, which is a major difficulty in the domain of FND. This comparison can also help to create more effective prompting tactics and guide future research on LLM-based FND systems.

Therefore, this study aims to analyze and compare the performance difference between using zero-shot and few-shot prompting style in FND. While zero-shot and few-shot prompting have been studied in various NLP tasks, this study applies these prompting styles to a multi-class fake news detection task using two benchmark datasets. Comparing zero-shot with two-shot prompting in this specific context reveals subtle nuances in LLM behavior when handling ambiguous and stylistically diverse fake news. This study employs a pre-trained LLM to detect fake news and approach it as a text classification problem. The goal is to elicit the inherent knowledge contained within the pre-trained LLM, which is a residual of its extensive training procedure, and determine if this knowledge is veracious enough to effectively aid the LLM's capability for FND. Therefore, the study follows the concepts outlined in [11], expecting that increasing the number of examples provided in the input prompt enhances the model's rationalization capabilities and improve classification performance. However, while [11] focuses on answering Theory-of-Mind questions, this study focuses the LLM on determining the truthfulness of news content.

2.0 RESEARCH METHODOLOGY

As discussed, the study aims to investigate whether differences in zero-shot or few-shot prompting styles lead to noticeable differences in LLM performance in the domain of FND. The study involves collecting LLM responses from a pair of benchmark datasets related to FND. The response collection is conducted in two settings: first, using a zero-shot prompting method, and second, using a two-shot prompting method. The scope of the study is as follows:

- Dataset Used: PolitiFact and Liar
- LLM Used: GPT-3.5-Turbo
- Prompting Style Tested: Zero-Shot & Two-Shot
- Metric Used for Evaluation: Accuracy, Precision, Recall & F1-Score

The study is conducted in several phases. The research methodology for the study is divided into the following phases:

- Phase 1: Dataset Preparation
- Phase 2: LLM Response Collection
- Phase 3: LLM Response Encoding
- Phase 4: Evaluation

The following sections of the study will elaborate on each phase involved in its methodology.

2.1 Phase 1: Dataset Preparation

The first phase is the data preparation phase. This phase involves obtaining the dataset that is used in the study. The main criterion for choosing the dataset is based on their popularity as benchmark datasets for FND tasks. Therefore, the dataset used for the study is PolitiFact [20] and Liar [21]. These datasets are popular datasets used primarily for FND in NLP. The details regarding these 2 datasets are outlined in Table 1 and Table 2.

Table 1: The benchmark datasets used in the study

Authors	**Dataset**	**Description**	**Number of Instances**
[20]	PolitiFact	A dataset of high-quality fact-check from the PolitiFact website.	21,152
[21]	Liar	A collection of 12.8K manually labelled short statements in various contexts obtained from the PolitiFact website.	12,836

Table 2: Features in the PolitiFact and Liar dataset

Dataset	Feature	Description
PolitiFact	Verdict	The verdict of the fact checks in one of 6 categories: true, mostly-true, half-true, mostly-false, false, pants-fire
	Statement originator	The person who made the statement.
	Statement	The fact-checked statement.
	Statement date	The date the statement was made.
	Statement source	The source of the statement.
	Fact checker	The person who fact-checked the statement.
	Factcheck date	The date when the fact-checked article was published.
	Factcheck analysis link	The link to the fact-checked analysis article.
Liar	Statement ID	A unique identifier for the statement.
	Label	Label in one of 6 categories: True, Mostly-true, Half-true, Barely-true, False, Pants-fire.
	Statement	The statement.
	Subject(s)	The subject(s) of the statement.
	Speaker	The speaker.
	Speaker's job title	The speaker's job title.
	State information	The state information.

	Party affiliation	The party affiliation.
	History of barely true counts	History of barely true counts (including the current statement).
	History of false counts	History of false counts (including the current statement).
	History of half true counts	History of half true counts (including the current statement).
	History of mostly true counts	History of mostly true counts (including the current statement).
	History of ‘pants on fire’ counts	History of ‘pants on fire’ counts (including the current statement).

While the dataset contains a number of features, the only features that are used for the purposes of this study is the ‘statement’ feature from the PolitiFact and Liar dataset. The GPT-3.5-Turbo model specializes in NLP tasks and does not require the user to vectorize their text for the model to process or comprehend it. This means the user will be able to provide the input prompt in natural language as the text processing method is done server-side when the GPT-3.5-Turbo model is queried via OpenAI’s official Application Programming Interface (API). Furthermore, providing the statement feature of PolitiFact and Liar to GPT-3.5-Turbo allows it to analyse the style present in the text, which is one of the most common perspectives to FND [22]. The GPT-series model of which GPT-3.5-Turbo belongs also demonstrated successful style detection and transfer capabilities [8]. Since LLMs are also able to demonstrate possession of factual knowledge, this study also incorporates a knowledge-based perspective to FND [22].

After obtaining the datasets, not all instances are used in the LLM response collection phase. Only a specific portion of each dataset is used for this phase. These portions are referred to as the validation sets. Figure 1 visualizes the number of data points present in the validation sets of each dataset used in the study compared to the original number of instances. A major reason for using a smaller sample of the datasets is to save time, reduce API costs, and preserve computing resources.

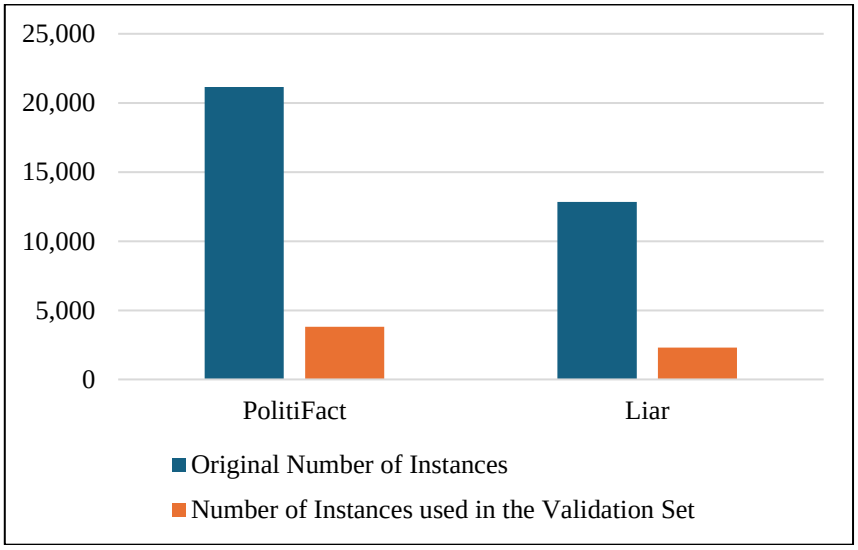

Fig. 1: Number of instances used in the validation set.

2.2 Phase 2: LLM Response Collection

The second phase is the LLM Response Collection phase. This phase involves creating 4 separate prompts that is used to collect LLM responses for the PolitiFact and Liar Dataset in both zero-shot and two-shot settings. The four prompts are detailed in Table 3.

Table 3. The four prompts used in the study.

Dataset	Category	Prompt

PolitiFact	Zero-shot	Please respond with only one word. Based on your knowledge, classify the following statement into one of the following 6 classes: True, Mostly-true, Half-true, False, Mostly-false, Pants-fire. (Statement Placeholder)
	Two-shot	1. Statement: "Doctors and nurses who administer the coronavirus vaccine can be "tried as war criminals."" Response: Pants-fire 2. Statement: "All the billionaires in America, their net worth combined … has increased by $800 billion" during the pandemic." Response: Mostly-true Based on your knowledge, please classify the following statement into one of the following categories: True, Mostly-true, Half-true, False, Mostly-false, or Pants-fire. Provide a one-word response: (Statement Placeholder)
Liar	Zero-shot	Please respond with only one word. Based on your knowledge, classify the following statement into one of the following 6 classes True, Mostly-true, Half-true, Barely-true, False, or Pants-fire. (Statement Placeholder)
	Two-shot	1. Statement: "Doctors and nurses who administer the coronavirus vaccine can be "tried as war criminals."" Response: Pants-fire 2. Statement: "All the billionaires in America, their net worth combined … has increased by $800 billion" during the pandemic." Response: Mostly-true Based on your knowledge, please classify the following statement into one of the following categories: True, Mostly-true, Half-true, Barely-true, False, or Pants-fire. Provide a one-word response: (Statement Placeholder)

As seen in Table 3, the zero-shot prompts are not provided with any examples, while the two-shot prompts include two examples. For the two-shot prompts, both examples are identical in the PolitiFact and Liar datasets, but the prompts ask the LLM to categorize the statements into their respective label sets, similar to the zero-shot prompts.

After preparing the prompts, a Python script was developed to make API requests to the OpenAI API to collect responses using the GPT-3.5-Turbo model. The script replaces the statement placeholder with each statement and requests a response from the GPT-3.5-Turbo model. The parameter for the maximum number of tokens returned by GPT-3.5-Turbo is set to 5 to ensure the responses are constrained to a single word. The requests are done in several batches and spaced out over several hours to respect OpenAI's API rate limit. In total, 12,234 API requests were made, and 12,234 GPT-3.5-Turbo responses were collected. The collected responses are then stored for use in the next phase of the study.

2.3 Phase 3: LLM Response Encoding

The third phase of the study is LLM response encoding. This phase involves converting the responses generated by GPT-3.5-Turbo into discrete numerical labels to perform more effective analysis. The generative nature of LLMs makes them difficult to use for discriminative tasks like FND. Therefore, specific steps must be taken to convert the text generated by GPT-3.5-Turbo into numerical labels. This allows us to compare the converted labels with the true labels from the PolitiFact and Liar datasets to properly evaluate the LLM's performance.

However, certain steps must be taken first to clean the data of unusable responses. In this study, an unusable response is defined as one where the response generated by GPT-3.5-Turbo does not belong to exactly one of the classes specified in the input prompt. Even with an input prompt specifically requesting a one-word reply categorized into six distinct categories, GPT-3.5-Turbo sometimes fails to follow these instructions, as shown in Table 4. These responses can sometimes be ambiguous or synonymous with the labels specified in the input prompt. However, since their total number is negligible (accounting for only 0.29% of the generated responses in the zero-shot setting and 0.48% in the two-shot setting), they are marked for automated removal later in the encoding process.

Table 4. Sample of unusable responses generated by GPT-3.5-Turbo in this study.

Statement	Response
Voted the 'best specialty plate in America'	Unknown.
Wisconsin "is on pace to double the number of layoffs" this year.	Uncertain
The Supreme Court's views "are radically out of step with public opinion" regarding its decision to legalize same-sex marriage nationwide.	Subjective
Rep. Carol Shea-Porter "votes with Nancy Pelosi's Democrats 95 percent of the time," but Frank Guinta "will take on both parties" and has "independent New Hampshire values.	Misleading
I represent the fourth-poorest (congressional) district" in the nation.	Unable to answer without additional[1]

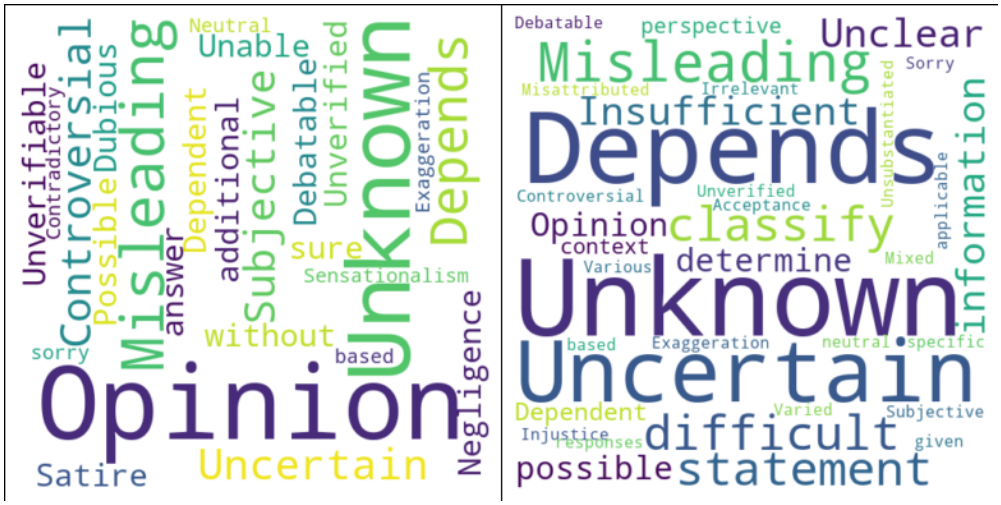

Fig. 2: Wordcloud of unusable responses in the zero-shot setting (left) and two-shot setting (right)

In the encoding process, a simple ruleset is used for automatically encoding the responses generated by GPT-3.5-Turbo. As mentioned earlier, the maximum number of tokens used by GPT-3.5-Turbo in generating its responses is intentionally limited to 5. This systematically ensures that responses consist of a single word only. Figure 2 shows two word clouds generated from a combination of all unusable responses in both zero-shot and two-shot settings. As

[1] Some responses are deliberately shortened because the max_tokens parameter is set to 5. This means the API will only return the first 5 tokens.

seen in Figure 2 and by the last example in Table 4, the intentional limiting of output tokens can produce incomplete phrases. Figure 3 visualizes the ruleset used for encoding the GPT-3.5-Turbo responses in this study.

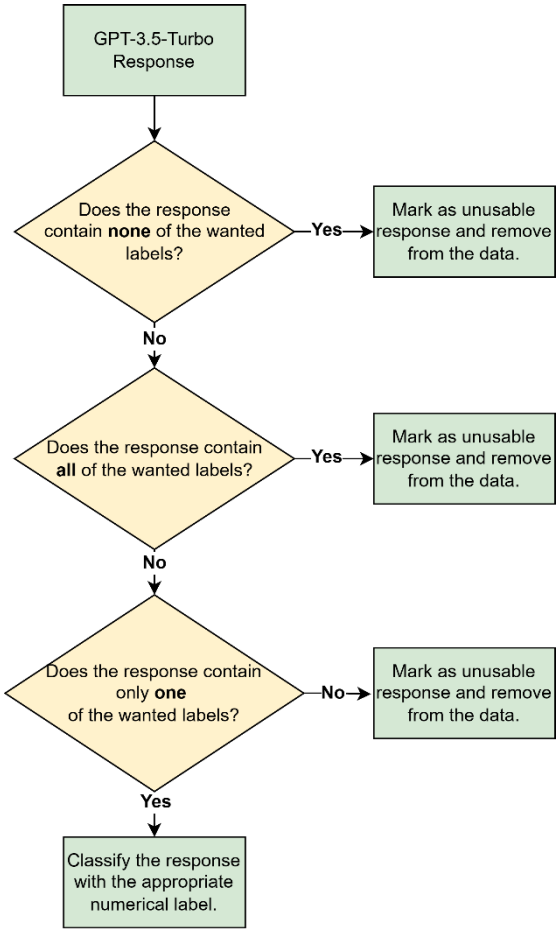

Fig. 3: The ruleset used for encoding the GPT-3.5-Turbo responses.

The encoded response is then recorded for further evaluation and analysis.

2.4 Phase 4: Evaluation

Evaluation is the final phase of the research methodology. It involves performing comparative analysis on the performance of the GPT-3.5-Turbo model in zero-shot and two-shot settings for the PolitiFact and Liar dataset. The four main metrics used to evaluate the performance of the LLM is accuracy, precision, recall and F1-score. Accuracy is a popular metric used in classification tasks and measures the proportion of correctly classified instances among the total number of instances evaluated. Another metric used in the study is the F1-score, which is the harmonic mean of precision and recall. The F1-score provides a balance between these two metrics, and is defined as follows:

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{1}$$

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

Where TP = True Positives (correctly predicted fake news), FP = False Positives (real news incorrectly predicted as fake news), and FN = False Negatives (fake news incorrectly predicted as real news). However, since the study

involves a multi-class aspect of FND, the F1-score calculation in this study involves calculating the score for each class individually and then taking the arithmetic mean of these class-wise F1 scores.

The F1-score is important because it provides a balanced measure of both precision and recall. A high F1-score ensures that the LLM can both accurately identify fake news and avoid false positives, which is important for maintaining public trust and preventing misinformation. Figure 4 illustrates an overview of the research methodology for this study discussed so far.

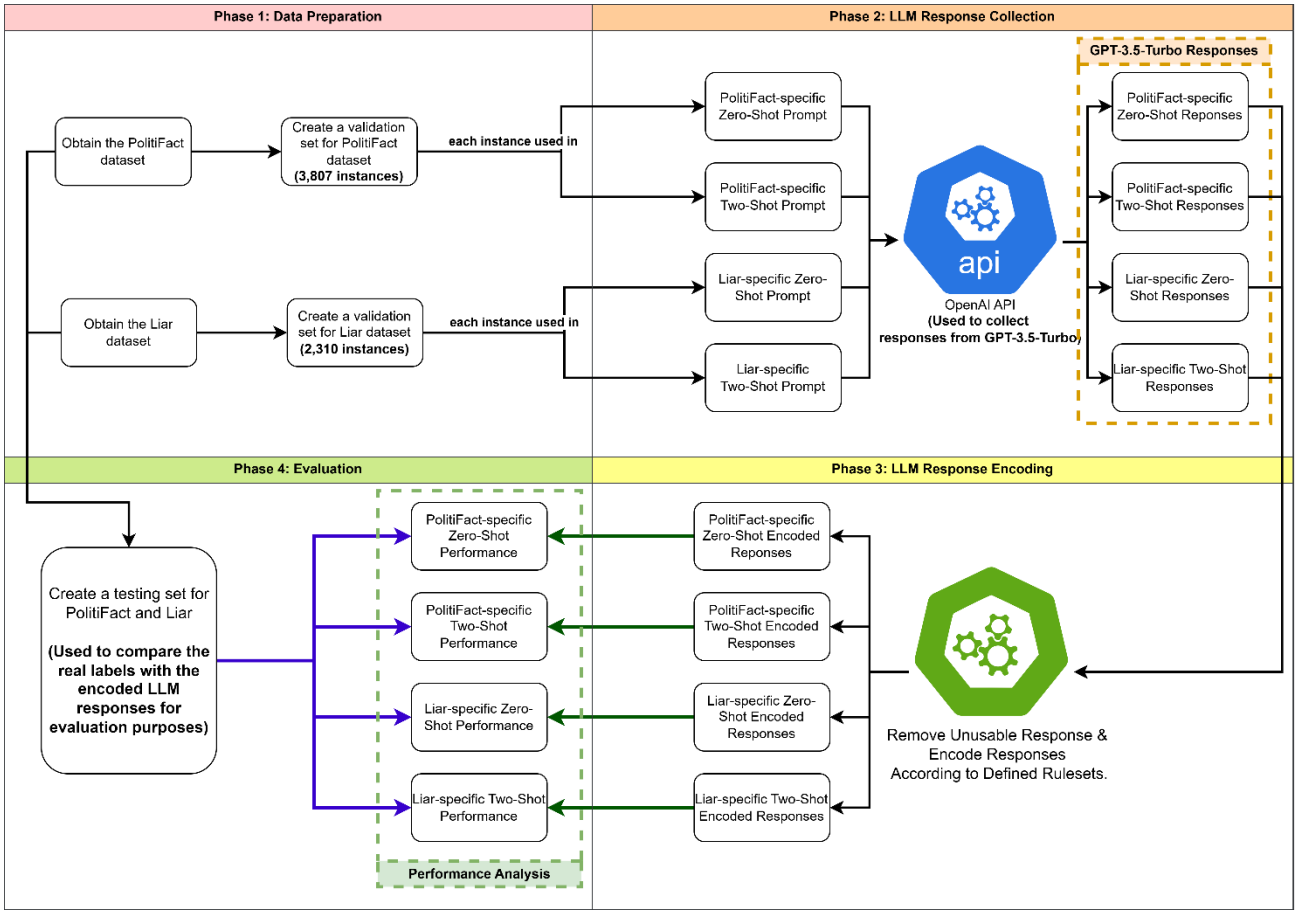

Fig. 4: An illustrative overview of the research methodology for this study

3.0 RESULTS

In the final phase of the study, a thorough performance analysis was conducted on the recorded results of the GPT-3.5-Turbo models in both zero-shot and two-shot settings. The evaluation focused on key metrics, namely accuracy, precision, recall, and F1-score, to assess the models' effectiveness in classifying data accurately. As stated earlier, accuracy measures the proportion of correctly classified instances, while the F1-score provides a balanced measure of precision and recall. Table 5 presents the experimental results across all settings, highlighting the comparative performance of the models. The findings highlight the importance of these metrics in understanding the strengths and weaknesses of the models in different scenarios.

Table 5. Performance of GPT-3.5-Turbo on PolitiFact and Liar in zero-shot and two-shot setting.

Dataset	Category	Accuracy	Precision	Recall	F1-Score
PolitiFact	Zero-shot	0.286	0.256	0.222	0.149
	Two-shot	0.293	0.271	0.255	0.232
Liar	Zero-shot	0.220	0.271	0.209	0.126
	Two-shot	0.226	0.240	0.218	0.199

4.0 DISCUSSION

4.1 Accuracy Analysis

According to Table 5, the accuracy of the LLM was found to be relatively low across both datasets and settings. However, a slight improvement was noted in the two-shot settings compared to the zero-shot settings. Specifically, the accuracy for the PolitiFact dataset increased from 0.286 in the zero-shot setting to 0.293 in the two-shot setting, while the accuracy for the Liar dataset increased from 0.220 to 0.226. These results denote that providing a few examples helps the model to some extent such as demonstrated in [11], but in the domain of FND, particularly when utilizing the PolitiFact and Liar dataset, the overall improvement is marginal. This suggests that the LLM struggles with the complexity of the classification tasks, even with additional context. The LLM also struggles with the style-based nuances required to differentiate between the six categories of fake news present in both datasets. Furthermore, it also suggests that the GPT-3.5-Turbo is unable to utilize any inherent factual knowledge to rationalize its predictions. This may be due to the ambiguous nature of the statements contained in the PolitiFact and Liar dataset, or it may also be due to the parameters set for the GPT-3.5-Turbo model during the LLM Response Collection Phase.

4.2 Precision, Recall, and F1-Score Analysis

Precision, recall, and F1-score metrics also demonstrated modest improvements in the two-shot settings for both datasets. For the PolitiFact dataset, precision increased from 0.256 to 0.271, recall from 0.222 to 0.255, and F1-score from 0.149 to 0.232. The Liar dataset also experienced an increase in recall from 0.209 to 0.218, and F1-score from 0.126 to 0.199. These slight enhancements across the metrics show that the LLM benefits from additional examples but still faces some challenges. The incremental gains suggest that while two-shot learning provides some additional context, it is insufficient to significantly improve the LLM's performance.

4.3 Comparative Analysis Across Datasets

Based on the performance of the two datasets, the GPT-3.5-Turbo model performed better on the PolitiFact dataset than on the Liar dataset in both settings. This is due to the LLM consistently achieving higher values for accuracy, precision, recall, and F1-score in the PolitiFact dataset. For example, the F1-score for PolitiFact in the two-shot setting is 0.232, whereas it is 0.199 in the Liar dataset. This disparity shows that the model may be more adept at handling the type of data or the specific structure of text present in the PolitiFact dataset. It also shows the variability in LLM performance across different datasets and how dataset-specific fine-tuning may be important.

4.4 Limitations and Future Implications

In addition to the challenges discussed earlier, human subjectivity in the labeling process of the original datasets, and the ambiguity and contextual dependence contained in the statements can further complicate classification, which leads to poor performance by the LLM due to its lack of these additional context. This is because PolitiFact and Liar are based on political statements, which may evoke certain biases from the human fact-checkers depending on their political alignment. These biases can introduce noise into the data, potentially hindering the model's ability to accurately generalize and classify ambiguous statements. Therefore, the findings of this study may not generalize well to non-political domains of misinformation, such as health or financial sectors. Addressing these challenges requires careful data curation and evaluation techniques. Beyond human biases, the low performance may also be influenced by the task's inherent subjectivity and the complexity of categorizing multi-class labels. These factors make this study a practical framework for evaluating the limits of ICL for LLMs in the context of fake news. The marginal performance gains offer insights into both the limitations and potential of GPT-3.5-Turbo in handling nuanced classification tasks. However, these improvements were not subjected to statistical significance testing due to the high cost of acquiring a larger sample size. Consequently, while the marginal gains can be informative, the lack of thorough statistical validation is a key consideration for future work.

The minor increases in performance from zero-shot to two-shot settings suggest that the GPT-3.5-Turbo model would benefit from more examples. However, the overall effectiveness of the LLM remains limited in the context of classifying fake news. While the gains in F1-score are noticeable, they are not significant and indicate the difficulty for the LLM to balance between accuracy and recall. In the context of multi-class FND, where both false positives and false negatives can have serious consequences, these measures are critically important. The low F1-scores, especially in the zero-shot scenarios (0.149 and 0.126 for PolitiFact and Liar respectively) illustrate the difficulty of achieving this balance. Additionally, the experiment uses only two datasets, both focused on political statements, which limits the generalizability of the results. This emphasizes the need for further optimization and potentially more complex methods to enhance the model's performance.

5.0 CONCLUSION

In conclusion, this study used the PolitiFact and Liar datasets to examine the effects of zero-shot and two-shot prompting approaches on the GPT-3.5-Turbo model's performance in the domain of FND. The study found that although two-shot prompting outperforms zero-shot prompting in terms of recall, accuracy, precision and F1-score, the model's total performance is still limited. The accuracy of the model rose very little in the two-shot setting for PolitiFact (from 0.286 to 0.293) and Liar (from 0.220 to 0.226). This suggests that providing examples helps the model to some extent but it still faces challenges with the complexity of the classification tasks inherent in these two datasets.

Additionally, the performance metrics demonstrated marginal gains in accuracy, recall, and F1-score for both datasets in the two-shot scenarios. Liar's F1-score rose from 0.126 to 0.199, while PolitiFact's F1-score went from 0.149 to 0.232. However, these little improvements imply that the extra information that two-shot learning offers is insufficient to significantly enhance the model's performance. The low F1-scores emphasize the need for more optimization and highlight the challenge of achieving a balanced measure of precision and recall in multi-class FND. Overall, these findings reveal that while the GPT-3.5-Turbo model has some potential, it is still not entirely prepared to handle multi-class FND. The performance increases due to more context (two-shot learning) are encouraging, but it is not adequate for serious real-world applications. More research and development are required to rectify this inadequacy.

Future work can explore more advanced prompting methods, such as Chain-of-Thought and Self-Consistency. Furthermore, increasing the number of shots (for example, five-shot or ten-shot) may provide deeper insights into model behavior and clarify the point of diminishing returns. Next, conducting a comprehensive error analysis to understand misclassifications and refining prompts to address ambiguous language and subtle stylistic cues would also be beneficial. Investigating whether LLMs exhibit biases toward simpler labels (for instance, "True," "False") over compound labels (for example, "Mostly-true") can be achieved through detailed F1-score analyses of individual classes. Increasing dataset variety by using additional benchmark datasets like FakeNewsNet or LIAR-PLUS could further improve performance and enhance robustness. Additionally, exploring more advanced GPT-series variants (GPT-4, GPT-4o, or OpenAI o1) and comparing them with both other LLMs (LLaMA, PaLM, Claude) and traditional ML methods (LR, RF, XGB) would contextualize the strengths and limitations of GPT-3.5-Turbo. A thorough cost-benefit analysis is also important to evaluate feasibility at scale and incorporating additional evaluation metrics such as the Matthews Correlation Coefficient (MCC) and Area Under Curve (AUC) would provide a more robust performance assessment.

ACKNOWLEDGEMENT
This study was supported by the Fundamental Research Grant Scheme (FRGS/1/2021/ICT02/UKM/02/1) from the Ministry of Higher Education, Malaysia.

Conflicts of Interest
The authors declare that they have no conflicts of interest.

REFERENCES

[1] S. Koka, A. Vuong, and A. Kataria, "Evaluating the Efficacy of Large Language Models in Detecting Fake News: A Comparative Analysis," Jun. 2024, Accessed: Jun. 19, 2024. [Online]. Available: https://arxiv.org/abs/2406.06584v1

[2] S. Loomba, A. de Figueiredo, S. J. Piatek, K. de Graaf, and H. J. Larson, "Measuring the impact of COVID-19 vaccine misinformation on vaccination intent in the UK and USA," *Nature Human Behaviour 2021 5:3*, vol. 5, no. 3, pp. 337–348, Feb. 2021, doi: 10.1038/s41562-021-01056-1

[3] S. K. W. Chu, S. D. Ravana, S. S. W. Mok, and R. C. H. Chan, "Behavior, perceptions and learning experience of undergraduates using social technologies during internship," *Educational Technology Research and Development*, vol. 67, no. 4, pp. 881–906, Aug. 2019, doi: 10.1007/S11423-018-9638-2/METRICS

[4] J. Alghamdi, Y. Lin, and S. Luo, "A Comparative Study of Machine Learning and Deep Learning Techniques for Fake News Detection," *Inf.*, vol. 13, no. 12, Dec. 2022, doi: 10.3390/INFO13120576

[5] M. Sami and A. B. M. Shawkat Ali, "Machine Learning Algorithms Performance Investigation in Fake News Detection," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 13864 LNCS, pp. 95–110, 2023, doi: 10.1007/978-981-99-2233-8_7

[6] S. A. Alameri and M. Mohd, “Comparison of Fake News Detection using Machine Learning and Deep Learning Techniques,” *2021 3rd International Cyber Resilience Conference, CRC 2021*, Jan. 2021, doi: 10.1109/CRC50527.2021.9392458

[7] L. Ouyang *et al.*, “Training language models to follow instructions with human feedback,” *ArXiv*, vol. abs/2203.02155, 2022

[8] T. B. Brown *et al.*, “Language Models are Few-Shot Learners,” *Adv Neural Inf Process Syst*, vol. 2020-December, May 2020, Accessed: May 15, 2023. [Online]. Available: https://arxiv.org/abs/2005.14165v4

[9] OpenAI, “GPT-4 Technical Report,” *ArXiv*, vol. abs/2303.08774, 2023

[10] Y. Huang and L. Sun, “FakeGPT: Fake News Generation, Explanation and Detection of Large Language Models,” Oct. 2023, Accessed: Jun. 19, 2024. [Online]. Available: https://arxiv.org/abs/2310.05046v2

[11] S. R. Moghaddam and C. J. Honey, “Boosting Theory-of-Mind Performance in Large Language Models via Prompting,” Apr. 2023, Accessed: Mar. 16, 2024. [Online]. Available: https://arxiv.org/abs/2304.11490v3

[12] Y. Li, “A Practical Survey on Zero-shot Prompt Design for In-context Learning,” *International Conference Recent Advances in Natural Language Processing, RANLP*, pp. 641–647, Sep. 2023, doi: 10.26615/978-954-452-092-2_069

[13] K. Taguchi, Y. Gu, and K. Sakurai, “The Impact of Prompts on Zero-Shot Detection of AI-Generated Text,” Mar. 2024, Accessed: Jun. 19, 2024. [Online]. Available: https://arxiv.org/abs/2403.20127v1

[14] W. Gao, M. Ni, H. Deng, X. Zhu, P. Zeng, and X. Hu, “Few-shot fake news detection via prompt-based tuning,” *Journal of Intelligent & Fuzzy Systems*, vol. 44, no. 6, pp. 9933–9942, Jun. 2023, doi: 10.3233/JIFS-221647

[15] A. Shah, S. D. Ravana, S. Hamid, and M. A. Ismail, “Web credibility assessment: Affecting factors and assessment techniques,” *Information Research*, vol. 20, Mar. 2015

[16] B. Hu *et al.*, “Bad Actor, Good Advisor: Exploring the Role of Large Language Models in Fake News Detection,” *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 20, pp. 22105–22113, Sep. 2023, doi: 10.1609/aaai.v38i20.30214

[17] J. Wu, S. Li, A. Deng, M. Xiong, and B. Hooi, “Prompt-and-Align: Prompt-Based Social Alignment for Few-Shot Fake News Detection,” *International Conference on Information and Knowledge Management, Proceedings*, pp. 2726–2736, Sep. 2023, doi: 10.1145/3583780.3615015

[18] E. Mitchell, Y. Lee, A. Khazatsky, C. D. Manning, and C. Finn, “DetectGPT: Zero-Shot Machine-Generated Text Detection using Probability Curvature,” *Proc Mach Learn Res*, vol. 202, pp. 24950–24962, Jan. 2023, Accessed: Jun. 19, 2024. [Online]. Available: https://arxiv.org/abs/2301.11305v2

[19] Q. Li and W. Zhou, “Connecting the Dots Between Fact Verification and Fake News Detection,” *COLING 2020 - 28th International Conference on Computational Linguistics, Proceedings of the Conference*, pp. 1820–1825, Oct. 2020, doi: 10.18653/v1/2020.coling-main.165.

[20] Misra and Rishabh, “Politifact Fact Check Dataset,” Sep. 2022. doi: 10.13140/RG.2.2.29923.22566.

[21] W. Y. Wang, “‘Liar, liar pants on fire’: A new benchmark dataset for fake news detection,” *ACL 2017 - 55th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)*, vol. 2, pp. 422–426, 2017, doi: 10.18653/V1/P17-2067.

[22] X. Zhou and R. Zafarani, “A Survey of Fake News,” *ACM Computing Surveys (CSUR)*, vol. 53, no. 5, Sep. 2020, doi: 10.1145/3395046.