Quest Journals
Journal of Software Engineering and Simulation
Volume 11 ~ Issue 8 (August 2025) pp: 28-42
ISSN(Online) :2321-3795 ISSN (Print):2321-3809
www.questjournals.org

Research Paper

A Dual-Path Approach to Fake News Detection Based on Ensemble Learning and Fine-Tuned Large Language Models

Jiayi Tang
Jinling High School

ABSTRACT: *The rapid proliferation of misinformation in today's digital ecosystem poses threats to public's trust on institutions and informed decision-making. To mitigate these challenges, the development of efficient and automated fake news detection systems has become increasingly essential. This study presents a dual-path framework, FIND, an ensembled model, combines Random Forest classifiers, fine-tuned BERT models, and prompt-engineered LLMs to address mainstream detection tasks with high accuracy. Additionally, we fine-tune LLMs to adapt to long-tail news (ie news with uncommon topics) distributions, enhancing model generalizability and scalability. Experimental evaluations on benchmark datasets confirm the robustness and effectiveness of the proposed method in handling complex fake news, proving its applicability in real-world scenarios.*

KEYWORDS: *fake news detection, machine learning, large language model, ensembled learning, fine-tuned LLM*

***R**eceived 08 Aug., 2025; **R**evised 19 Aug., 2025; **A**ccepted 21 Aug., 2025 © The author(s) 2025.*
P**ublished with open access at **www.questjournas.org

I. INTRODUCTION

The proliferation of fake news has become a critical issue in today's information ecosystem [1] influencing public opinion, inciting social unrest, and undermining trust in institutions. Allcott and Gentzkow [2] indicated that during the 2016 US presidential election, an overwhelming amount of fake news amassing over 37 million shares in a relatively

short three-month period before the election. Many Trump supporters was influenced by popular fake news stories that tended to favor Donald Trump over Hillary Clinton. According to NBC, teenagers from the Macedonian town of Veles earned at least $60,000 in six months by creating such fake news for millions on social media during that period [3].

The detrimental effects of fake news dissemination extend beyond the political realm, encompassing severe health-related consequences. Notably, there have been tragic instances in which online advertisements for experimental cancer treatments, erroneously perceived as credible medical information, have led to the untimely deaths of cancer patients. Similarly, false or misleading assertions regarding the COVID-19 virus have posed significant threats to public health, as individuals have been influenced to engage in risky behaviors, such as consuming harmful substances or neglecting social distancing guidelines. The COVID-19 pandemic has seen a marked increase in the spread of fake news, often characterized by sensational content that misleads individuals into believing in its purported efficacy [4]. Alarmingly, within a two-month period, the International Fact-Checking Network (IFCN) identified over 3,500 false claims related to COVID-19. For example, the propagation of misinformation suggesting unproven remedies or linking the virus to 5G technology has resulted in physical harm. Tragically, according to BBC [5], it is estimated that at least 800 individuals globally may have lost their lives in the first three months of 2020 due to false claims associated with the coronavirus.

As the volume of content disseminated online continues to grow, traditional fact-checking methods struggle to keep pace with the sheer scale and speed of false information. In response, automated approaches for detecting fake news have gained significant attention from the research community, leveraging advances in natural language processing (NLP) to enhance accuracy and efficiency [6]. Figure 1 shows the growing interest in fake news detection over the last decade, as extracted from Dimensions AI [7]. Recent developments in large language models (LLMs), such as GPT [8], BERT [9], and their successors [10], have revolutionized NLP by offering powerful capabilities in understanding and generating human language. These models, pre-trained on vast corpora and fine-tuned for specific tasks, demonstrate remarkable performance in various text-based applications, including sentiment analysis, machine translation, and text summarization. Importantly, LLMs also hold promise for the automated detection of fake news due to their ability to capture nuanced linguistic patterns and contextual information.

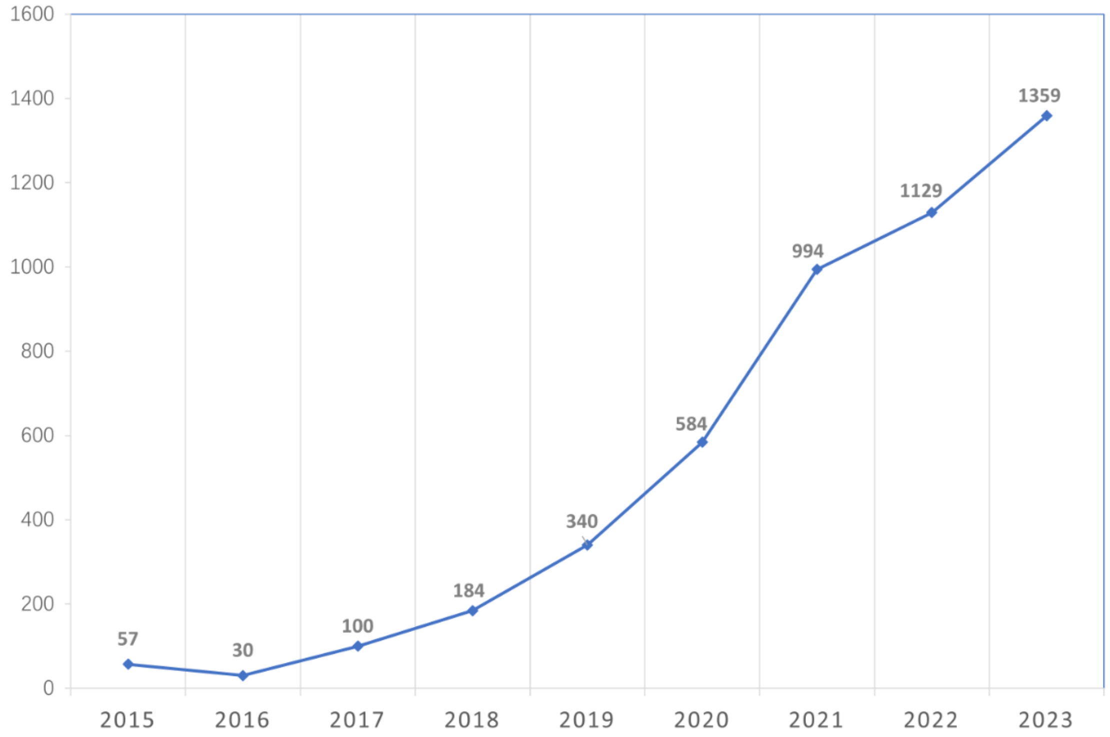


Figure 1: Publications on fake news detection

In this work, I provide a comprehensive review of existing machine learning and LLM based approaches for fake news detection and extensive experiments evaluating their effectiveness and corresponding drawbacks. Also, we propose a **F**ine-tuning and **I**n-context learning based LLM framework for fake **N**ews **D**etection (**FIND**) By combining Fine-tuned LLM and in-context learning techniques, FIND combines the strengths of both and outperforms them under the experiment settings. In the experiments, a new dataset containing statements collected from Twitter is used for their evaluation.

II. RELATED WORKS

Fake news detection is a binary task between real and fake items [11]. The challenge of fake news detection has attracted considerable attention across various domains, leading to the development of numerous methodologies and frameworks. Early approaches primarily relied on traditional machine learning techniques, employing handcrafted features such as linguistic cues, metadata, and user engagement metrics. These methods, while effective to some extent, often struggled to generalize across diverse datasets and evolving misinformation strategies.

With the advent of deep learning, researchers began to explore neural network architectures for fake news detection. Models such as convolutional neural networks [12] and recurrent neural networks [13] showed improved performance by leveraging learned representations of textual data. However, these models still faced limitations in capturing the complexities of language and context.

In recent years, large language models (LLMs) have emerged as a transformative force in the field of natural language processing. By pre-training on extensive datasets and employing attention mechanisms, LLMs such as BERT and GPT have demonstrated superior capabilities in understanding semantic relationships and contextual nuances. LLMs performed outstandingly on a variety of NLP tasks, including information extraction, text classification, and sentiment analysis [14] Several studies have begun to explore the application of LLMs specifically for fake news detection, focusing on their ability to discern factual accuracy through contextual embeddings and transfer learning techniques.

Few-Shot prompting provides prompts to specific tasks and several labeled examples [15]. By utilizing its pre-trained knowledge, the model can predict the veracity of news articles with only a few labeled examples [16]. Previous studies [17] [18] [19] used few-shot prompting for fake news detection.

Another critical advancement is the integration of external data sources in the LLM-based fake news detectors (e.g. through API integrations). Online LLM models have access to real-time information, making detection more accurate [20].

Fine-tuning BERT is another method utilized in fake news detection. The process minimizes the computational cost, since only the added task-specific layers are trained, while the original Transformer architecture remains largely unchanged [21].

However, the potential of LLMs is severely limited when external information is scarce [22]. Also, Hu et al. [23] found that current LLMs are not capable to substitute fine-tuned SLMs in fake news detection, but it can be good advisor for it by providing multi-perspective instructive rationales. Accordingly, many models, including DAFND [24], ARG [25], and FactAgent [26], est., use external data and internal knowledge of the model respectively to detect, then use another module to make the final decision.

III. METHODOLOGY

3.1 Logistic regression

Logistic regression (LR) is a classifier for decomposing a dataset where one or more independent variables determine a result [27]. It consists of a vector feature representation, a function of classification to compute the estimated class $\hat{y}$ with sigmoid function via P(y|x), and cross-entropy loss function for learning optimization. It returns y=0 or y=1 based on the probability. The function of logistic regression is

$$p(X, b, w) = \frac{1}{1 + e^{w \cdot X + b}}$$

Logistic regression is straightforward to implement and understand. The coefficients represent the change in the log-odds of the outcome for a one-unit change in the predictor, making it easy to interpret. It requires relatively little computational power compared to more complex models like neural networks or ensemble methods. In cases with a small number of features, logistic regression is less likely to overfit than more complex models, particularly when regularization techniques are applied. However, for more complex data with nonlinear relationships, it is hard for logistic regression to offer satisfactory performance.

3.1 Decision Tree

Decision tree is a non-parametric supervised learning algorithm. A decision tree works by splitting the data into subsets based on the value of input features, creating a tree-like structure where each internal node represents a decision based on a feature, each branch represents the outcome of that decision, and each leaf node represents a predicted output [28]. Typically, a decision tree begins with a root node that has no branches coming in. The internal nodes, sometimes referred to as decision nodes, receive the outgoing branches from the root node. Both node types perform assessments to create homogeneous subsets, represented by leaf nodes or terminal nodes, based on available features. All outcomes of the dataset are represented by the leaf nodes. Figure 2 below is a toy example of a decision tree for weather forecast.

Decision trees are easy to understand and interpret. The visual representation makes it straightforward to see how decisions are made, providing excellent explainability, which is valuable for interpret the results. Also, decision trees can model complex nonlinear relationships effectively, as they make decisions based on various feature thresholds.

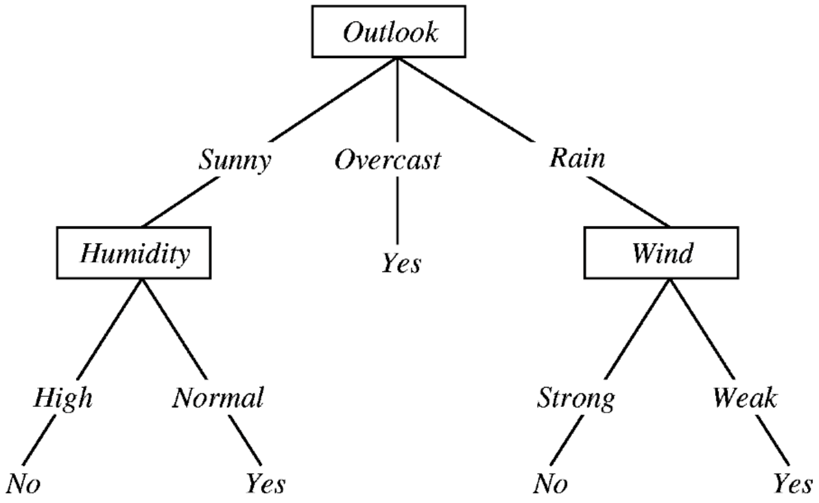


Figure 2: A example of decision tree

A random forest incorporates a combination of tree-structured classifiers

$$\{h(x, \Theta_k), k = 1,2, \ldots\}$$

where the Θ_k are distributed independent random vectors identically and each tree casts a unit vote for the most popular class at input x [29].

3.4 Random Forest

Random Forest is an ensemble learning technique primarily used for classification and regression tasks. It operates by constructing a multitude of decision trees during training time and outputting the mode of the classes (for classification) or the mean prediction (for regression) of the individual trees. The key advantage of Random Forest lies in its ability to mitigate overfitting, a common issue with single decision trees, by averaging the results of multiple trees, which leads to improved model accuracy and robustness [30].

The algorithm works by randomly sampling subsets of the training data, a process known as bootstrap aggregating or bagging. For each subset, a decision tree is constructed, and during the tree-building process, a random subset of features is considered for splitting at each node. This introduces additional randomness into the model and helps ensure that the trees are decorrelated, which enhances the overall performance of the ensemble [31].

Random Forest also provides insights into feature importance, allowing practitioners to understand which features are most influential in making predictions. This characteristic is particularly useful in applications where interpretability is essential [32]. Overall, Random Forest is widely adopted across various domains due to its flexibility, ease of use, and superior performance in a range of scenarios.

3.5 Fine-tuned BERT

BERT [33] is a pre-trained deep learning model designed for natural language processing tasks proposed by Google in 2018. BERT leverages the Transformer architecture, which allows it to process text in a bidirectional manner. It is pre-trained on a large corpus of text using two primary tasks: masked language modeling and next sentence prediction. BERT's architecture consists of multiple layers of encoders, each comprising self-attention mechanisms that enables BERT to capture intricate linguistic features and dependencies, making it particularly effective for a variety of downstream tasks such as sentiment analysis, question answering, and named entity recognition.

We use BERT as the base LLM for fine-tuning. The model contains an encoder with 12 Transformer blocks and 12 self-attention heads. It takes an input of a sequence of no more than 512 tokens and outputs the representation of the sequence with 768 dimensions. In order to adapt the BERT model to fake news detection task, the full parameter fine-tuning method is adopted by using a full-connected network as the classifier layer to aggregate the semantic information from the BERT output. The model architecture is given as Figure 3.

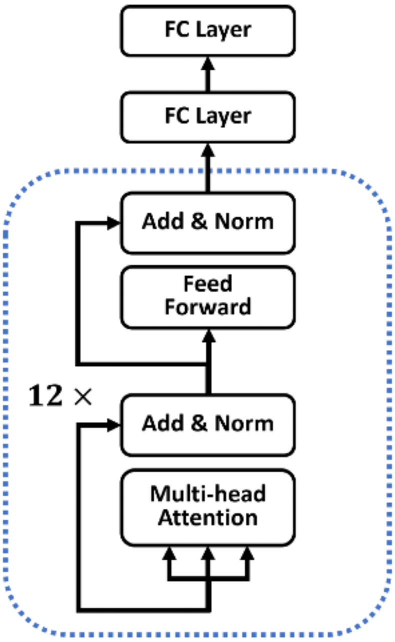


Figure 3: Fine-tuned BERT Model.

3.6 Ensemble Learning

Ensemble learning is a powerful machine learning paradigm that combines multiple individual models, often referred to as "base learners," to improve overall predictive performance. The central premise of ensemble methods is that by aggregating the predictions of diverse models, one can achieve better accuracy and robustness than any single model could provide [34]. This approach is particularly effective in mitigating the issues of overfitting and bias, allowing for improved generalization to unseen data.

We propose FIND in this part, an ensemble learning framework for effective fake news detection. FIND integrates three distinct classifiers: random forest model, fine-tuned BERT model, and prompted LLM. The structural differences among these three models are significant, allowing them to complement each other effectively. Two ways can be used to combine the outputs from these three classifiers. The first is to use logistic regression as the meta model and adjust each classifier's contribution based on its performance during the training process. The second is to use a weighted vote and decide the weights of models based on their performance on the validation set. The ensemble approach harnesses the strengths of each model, leading to a more robust and accurate system for detecting fake news.

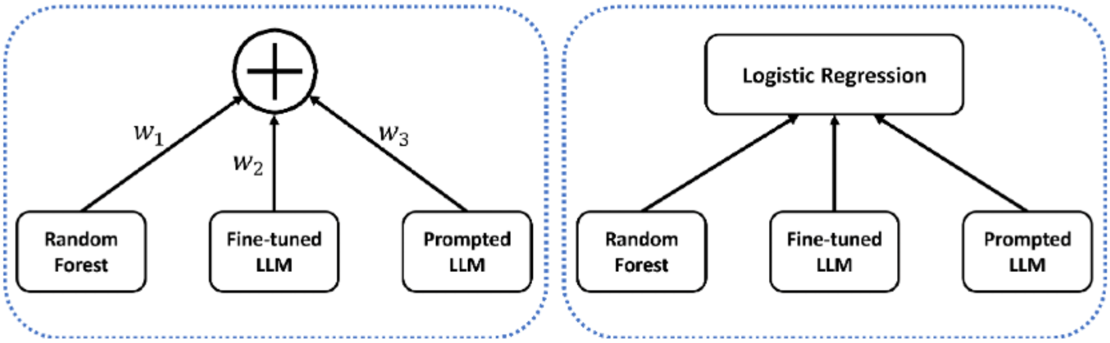


Figure 4: two variations of FIND model

3.8 In-context Learning Workflow

I implement a Coze workflow into AI-agent, which allows for adaptive and iterative processes.

In general, workflows replace linear interactions to improve LLMs' functionality [35]. It breaks tasks into sequential, goal-oriented steps that will be completed collaboratively by different agents that focus on specific task. Integrating external engines (e.g., APIs or databases) in the workflow allows it to fetch real-time data and perform beyond their original scope.

Similar to iterative learning, in order to refine future interactions, it can assess prior outputs for self-improvement [36].

3.8.1 Curb Hallucinations

Prompt engineering is critical for anchoring LLM's responses. By Simplifying complex tasks into more digestible components and conveying explicit instructions, those prompts channel the LLM towards more accurate and relevant outputs, which can significantly curtail the rates of hallucinations. Also, incorporating examples within prompts, a technique known as few-shot learning, provides the model with concrete references, clarifying the expected format and substance of the response. Simultaneously, when integrating LLMs into product designs, striking this balance between accuracy and computational efficiency [37]. In order to design effective and sustainable deployment of LLMs, instead of using the conventional fixed prompt, I utilized dynamics prompt generation.

3.8.2 Workflow

The workflow allows LLMs to break down complex problems into manageable sub-steps within a structured workflow, by collaborating each component with different functions [38]. The search engine collects relevant information from search engines to enable the LLM to check facts in the news. The prompt engine generates the prompt base on the news content and the search results dynamically instead of outputting a fixed prompt for all news. By doing so, the LLM receives the prompt tailed for different news and facts. The decision engine gets the output from the search engine and the prompt engine and employ the LLM to decide whether the news is fake.

Search engine This engine utilizes the API of Google search engine to search for relative information reported on social media platforms. Its input is the input news, while the output includes url, publisher's information, title, author, news content, source, time and thumbnail. This engine helps mitigate the hallucination issue of LLM by using external knowledge to cross-reference and verify the news claim.

Prompt engine The prompt engine is intended for generating prompts tailoring to the news inputs and the result returned by the search engine. The prompt's aspect includes Core Information Extraction, Sentiment Analysis, Propagation and Influence Check, Contextual and Source Evaluation.

Decision engine This engine is specifically crafted for news classification. determine the authenticity of the news. It uses information comes from internet and judge it base on the prompt generated by the prompt engine. By

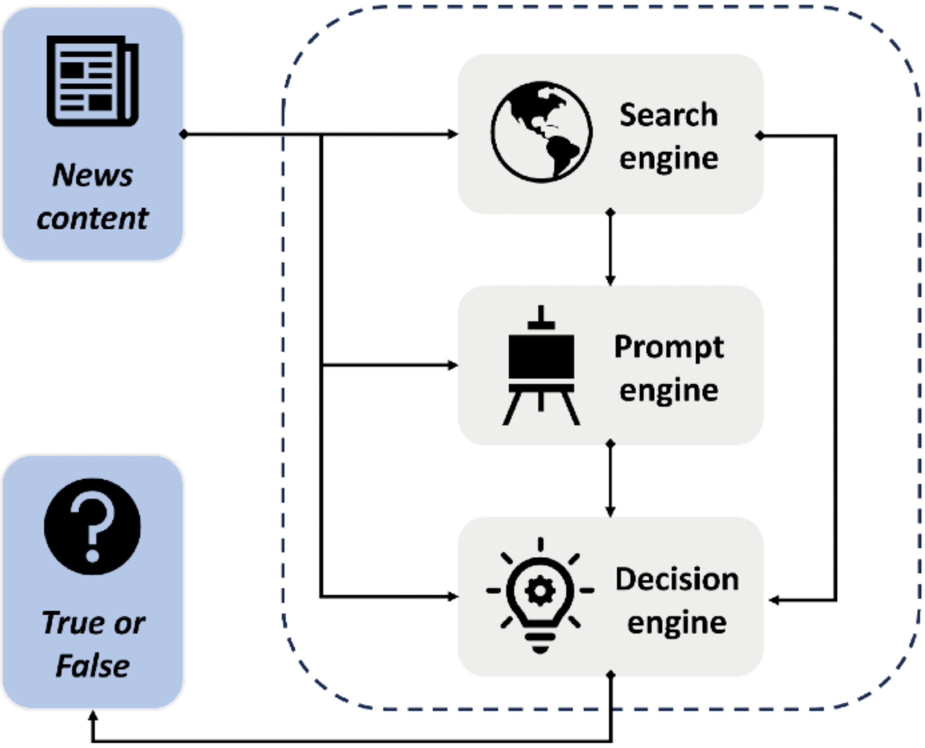
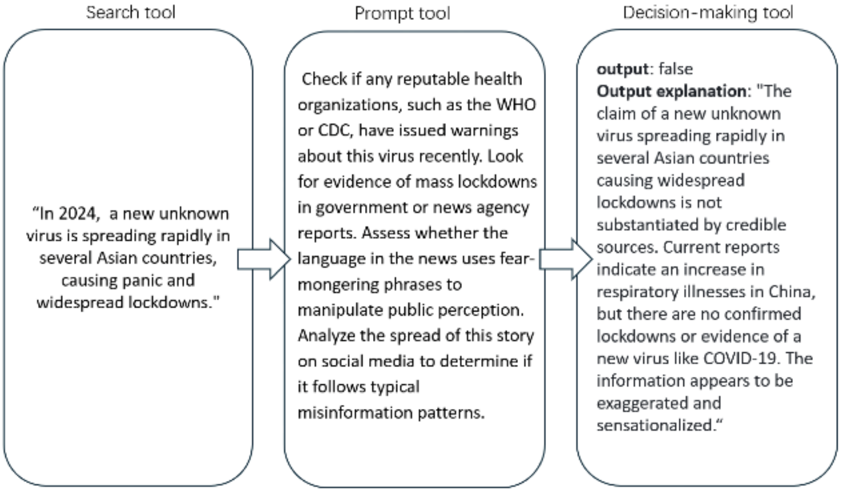


Figure 5: Illustration of the workflow



"In 2024, a new unknown virus is spreading rapidly in several Asian countries, causing panic and widesplead lockdowns."

Check if any reputable health organizations, such as the WHO or CDC, have issued warnings about this virus recently. Look for evidence of mass lockdowns in government or news agency reports. Assess whether the language in the news uses fear-mongering phrases to manipulate public perception. Analyze the spread of this story on social media to determine if it follows typical misinformation patterns.

output: false
Output explanation: "The claim of a new unknown virus spreading rapidly in several Asian countries causing widespread lockdowns is not substantiated by credible sources. Current reports indicate an increase in respiratory illnesses in China, but there are no confirmed lockdowns or evidence of a new virus like COVID-19. The information appears to be exaggerated and sensationalized."

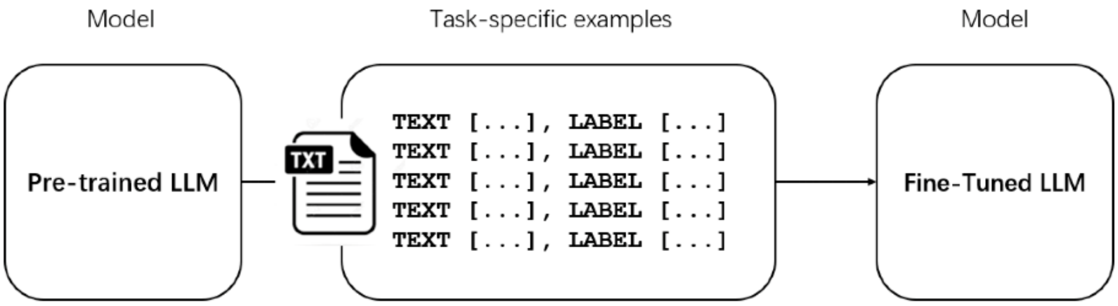
Figure 6: Workflow example

3.9 Fine-tuned Large Language Model

Beyond the FIND ensemble-based workflow, I fine-tuned a large language model using LoRA (Low-Rank Adaptation) to improve its ability to detect fake news [39].

The large model I used is from the Qwen series, released by Alibaba's Tongyi Laboratory, covering models from the lightweight Qwen-1.8B to the large-scale Qwen-72B. These models possess strong Chinese language processing capabilities as well as bilingual proficiency in Chinese and English, and are suitable for tasks such as dialogue generation, code understanding, and information extraction. The entire series has been fully open-sourced on Hugging Face. Among them, **Qwen2.5-7B-Instruct** is a variant specifically optimized for dialogue, incorporating alignment techniques during training to enhance safety, usefulness, and multi-turn conversation understanding [40]. On multiple Chinese NLP benchmark tests, these models perform exceptionally well under both zero-shot and few-shot settings, with strong scalability and robust safety evaluation foundations.

For small-scale downstream tasks (e.g., vertical-domain text generation, instruction understanding), we adopted the **LoRA** technique within the Parameter-Efficient Fine-Tuning (PEFT) framework. LoRA freezes the original model weights and trains only the low-rank matrices injected into each Transformer layer, thereby enabling rapid adaptation of model capabilities while significantly reducing GPU memory usage and computational costs. The PEFT/LoRA method can even be used to fine-tune large language models with nearly 7 billion parameters on Google Colab [41].



IV. EXPERIMENT

4.1 Dataset

In the experimental phase, I constructed a dataset comprising 17,080 news statements from Twitter, encompassing both verified and fabricated content. Using the Twitter API, I gathered a broad spectrum of political news and social media rumors. Each news item was evaluated and labeled by GossipCop, drawing on content analysis, social context, and the judgments of professional journalists and subject-matter experts. For each entry, I extracted essential metadata, including tweet ID, news URL, title, and full content. The dataset preserves a defined ratio of true to fabricated news, as presented below.

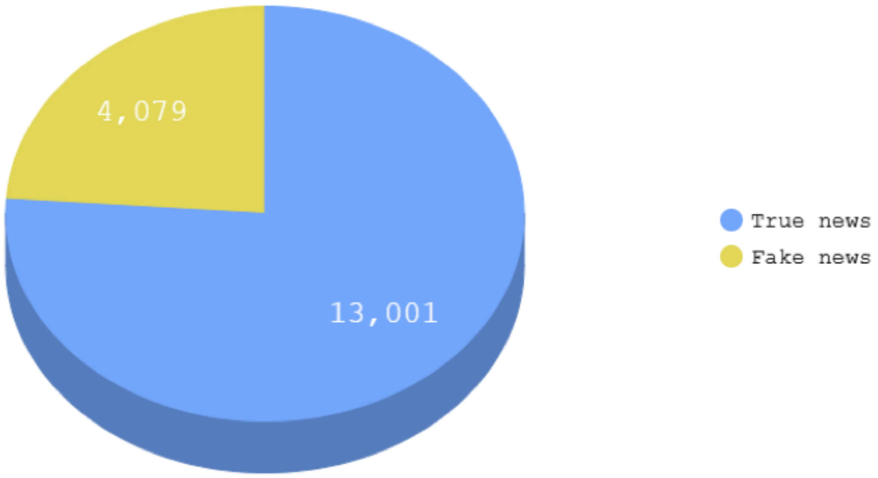
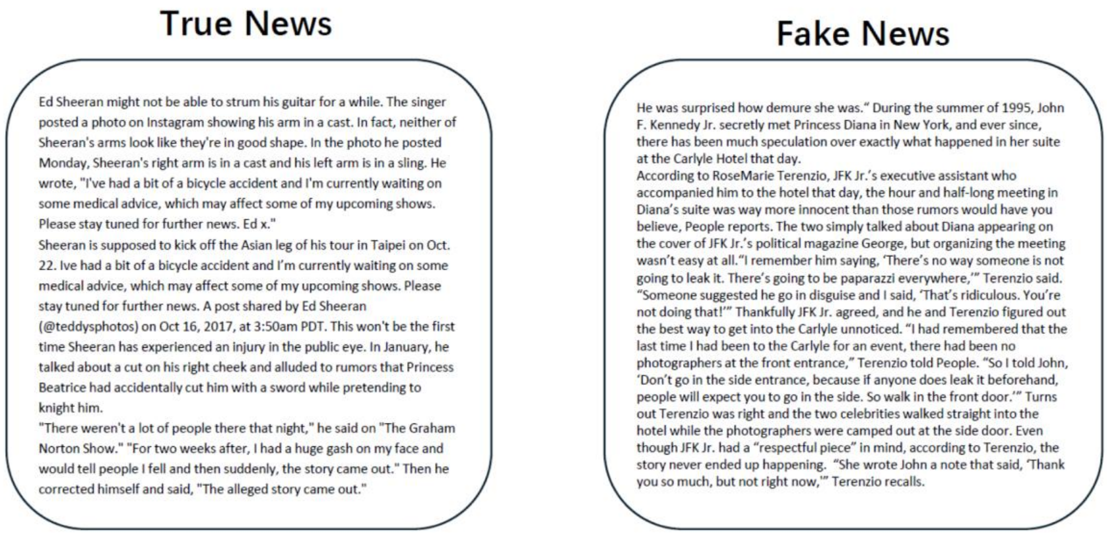


Figure 7 Ratio of true and fake news in the dataset

The dataset is mainly composed of posts which predominantly features content related to notable public figures and current political events. Here is the example of a true news and a fake news from the dataset as two examples in Figure.8.

The first one is about Ed Sheeran had an accident and is waiting on some medical advice, which is a true news that can be validated by Ed Sheeran's own posts. The second post is a tidbit of Princess Diana, implying that she had a tryst with the former American president John F. Kennedy, which is a rumor that has been exaggerated and maliciously speculated upon based on factual information according to existing information.



True News

Ed Sheeran might not be able to strum his guitar for a while. The singer posted a photo on Instagram showing his arm in a cast. In fact, neither of Sheeran's arms look like they're in good shape. In the photo he posted Monday, Sheeran's right arm is in a cast and his left arm is in a sling. He wrote, "I've had a bit of a bicycle accident and I'm currently waiting on some medical advice, which may affect some of my upcoming shows. Please stay tuned for further news. Ed x."

Sheeran is supposed to kick off the Asian leg of his tour in Taipei on Oct. 22. Ive had a bit of a bicycle accident and I'm currently waiting on some medical advice, which may affect some of my upcoming shows. Please stay tuned for further news. A post shared by Ed Sheeran (@teddysphotos) on Oct 16, 2017, at 3:50am PDT. This won't be the first time Sheeran has experienced an injury in the public eye. In January, he talked about a cut on his right cheek and alluded to rumors that Princess Beatrice had accidentally cut him with a sword while pretending to knight him.

"There weren't a lot of people there that night," he said on "The Graham Norton Show." "For two weeks after, I had a huge gash on my face and would tell people I fell and then suddenly, the story came out." Then he corrected himself and said, "The alleged story came out."

Fake News

He was surprised how demure she was." During the summer of 1995, John F. Kennedy Jr. secretly met Princess Diana in New York, and ever since, there has been much speculation over exactly what happened in her suite at the Carlyle Hotel that day.

According to RoseMarie Terenzio, JFK Jr.'s executive assistant who accompanied him to the hotel that day, the hour and half-long meeting in Diana's suite was way more innocent than those rumors would have you believe, People reports. The two simply talked about Diana appearing on the cover of JFK Jr.'s political magazine George, but organizing the meeting wasn't easy at all."I remember him saying, 'There's no way someone is not going to leak it. There's going to be paparazzi everywhere,'" Terenzio said. "Someone suggested he go in disguise and I said, 'That's ridiculous. You're not doing that!'" Thankfully JFK Jr. agreed, and he and Terenzio figured out the best way to get into the Carlyle unnoticed. "I had remembered that the last time I had been to the Carlyle for an event, there had been no photographers at the front entrance," Terenzio told People. "So I told John, 'Don't go in the side entrance, because if anyone does leak it beforehand, people will expect you to go in the side. So walk in the front door.'" Turns out Terenzio was right and the two celebrities walked straight into the hotel while the photographers were camped out at the side door. Even though JFK Jr. had a "respectful piece" in mind, according to Terenzio, the story never ended up happening. "She wrote John a note that said, 'Thank you so much, but not right now,'" Terenzio recalls.

Figure 8: Example true and fake news in the dataset

4.2 Data Pre-processing

Here is the method of pre-processing the data for detection.

1. Inserting a column "class" as target feature

2. By utilizing WordLemmatizer from nltk.stem, the code removes stopwords, lemmatizes the text which transforms all the word to the dictionary form (lemma). These two steps reduce noise and improve the model's efficiency [42].

3. NLP-Convert text to vectors using Tfidvectorizer. TF-IDF is a measure of originality of a word by comparing the number of times a word appears in a document with the number of documents the word appears in. TF-IDF vectors transform text into numerical features suitable for training models in text classification tasks. For example, different words' distinguishing TF-IDF scores can explicit the fake and true news' pattern of word usage. Also, using specific terms with high TF-IDF score that contains sentiment can help classify whether news' attitude is positive or negative [43].

4.Merge true and fake data frames and randomly shuffle the data frame. Then, split it training and testing dataset.

4.3 Case study of prompt choosing

In this part, we will discuss how to enhance prompt engineering for improved fake news detection, using the following three prompts as case studies.

A. Prompt 1: Expert Analysis

You are a news fact-checking expert who specializes in detecting fake news. Please read the following news content carefully and classify it according to its authenticity: Enter **'1'** to indicate that the news is a true report. Enter **'0'** to indicate that the news is false information. When making a judgment, please consider the following factors:

1. **Factual accuracy** – Are the facts in the news content supported by evidence?
2. **Source credibility** – Does the news come from a trustworthy media or organization?
3. **Logical consistency** – Is the news content logically self-consistent and free of obvious contradictions?
4. **Background verification** – Does the news event match known facts or historical records?

Based on the above factors, please provide only **one number** as an answer.

B. Prompt 2: Detailed Analysis Guide

As a senior news authenticity analyst, your task is to evaluate the authenticity of the following news content. Please classify according to the following criteria: Enter **'1'** to indicate that the news is a true report. Enter **'0'** to indicate that the news is false information.

Criteria:

1. **Fact check** – Can the events, data, or statements mentioned in the content be verified by reliable sources?

2. **Source evaluation** – Does the platform or source of the news release have a good reputation and reliability?
3. **Tone and wording** – Does the news use emotional, exaggerated, or inflammatory language?
4. **Consistency check** – Does the news content match the known information and background, and does it contain internal contradictions?
5. **Verification from multiple parties** – Has the news been reported and confirmed by other credible media?

Based on the above factors, please provide only **one number** as an answer.

C. Prompt 3: Comprehensive Factor Analysis

You are a professional fake news detection expert, responsible for analyzing and classifying the authenticity of news content. Please classify according to the following criteria: Enter **'1'** to indicate that the news is a true report. Enter **'0'** to indicate that the news is false information.

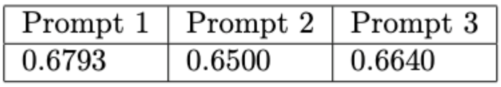
Criteria:

1. **Factual support** – Are the facts mentioned in the news supported by reliable evidence or sources?
2. **Source reliability** – Is the publisher of the news a recognized authoritative media or institution?
3. **Content logic** – Is the news content logically clear, and are there obvious contradictions or unreasonable points?
4. **Language style** – Is there any exaggerated, inflammatory, or biased language?

Based on the above factors, please provide only **one number** as an answer.

Choosing the best prompt}Randomly choose parts of the testing sample to test the prompt. The accuracy of each prompt is shown in the table.

Table 1: Accuracy of the three prompts



Prompt 1	Prompt 2	Prompt 3
0.6793	0.6500	0.6640

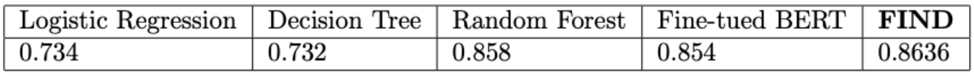
In this experiment, three distinct prompts were employed to assess the authenticity of news content through a large language model (LLM). Each prompt was designed to guide the LLM in evaluating various aspects of news articles, including factual accuracy, source credibility, logical consistency, and language style. The accuracy rates achieved by each prompt were as follows: Prompt 1 yielded an accuracy of 0.6793, Prompt 2 achieved an accuracy of 0.6500, and Prompt 3 reached an accuracy of 0.6640.

Prompt 1, which emphasized a comprehensive set of evaluation criteria, including factual accuracy and source credibility, demonstrated the highest accuracy among the three. This suggests that a more detailed and structured approach may enhance the model's ability to discern true from false information. Conversely, Prompt 2, despite its thoroughness, resulted in the lowest accuracy. The potential reasons for this could include the prompt's emphasis on emotional language and multiple source verification, which may have introduced ambiguity for the model.Prompt 3 performed moderately, indicating that while the criteria outlined were relevant, the phrasing and focus could have impacted the model's decision-making process. The variability in accuracy across the prompts highlights the importance of prompt design in eliciting optimal performance from LLMs in the context of fake news detection. As a result, prompt 1 is chosen as the best prompt for LLM to do fake news detection among these three prompts.

4.4 Model Evaluation

In this part, we test the models on our dataset. Five models are tested on the complete dataset: LR, random forest, fine-tuned BERT, prompted GPT, the ensemble model FIND. The accuracy of the models on the dataset is as follows.

Table 2: Accuracy of models



Logistic Regression	Decision Tree	Random Forest	Fine-tued BERT	**FIND**
0.734	0.732	0.858	0.854	0.8636

Analysis BERT-fine tuning and random forest classifier outperform logistic regression and decision tree, because they are more suitable for detecting data with diverse features. Feature diversity is the range and variability of input features. The true and fake news data have high diversity over linguistic constructs and sentiment tones. Logistic Regression and Decision tree classifier are simple models, performing well when dealing with linearly separable data and simple decision boundaries. When dealing with complex patterns like fake news, they have limited ability to extract complex patterns [44]. To that end, these models have a high possibility of underfitting or overfitting depending on the training size. On the other hand, BERT can capture nuanced, contextual information through its pre-trained Transformer architecture. Also, Random Forests combine various feature subsets and decision paths. It can handle noise and complexity effectively, which reduces the risk of over-fitting [45].

Detailed report of the ensemble model The classification report of the ensemble model is shown in the table

	Accuracy	Recall	f1 score	support
0	0.85	0.53	0.65	816
1	0.87	0.97	0.92	2600
weighted avg	0.86	0.86	0.85	3416

Table 3 Report of the ensemble model

Confusion Metrics In this study, we employ the confusion matrix to systematically analyze the performance of our proposed framework, providing insights into its strengths and areas for potential improvement. It provides a comprehensive breakdown of a model's performance by summarizing the correct and incorrect predictions made by the model across different classes. Each entry in the confusion matrix represents the number of instances classified in a particular way. The matrix is structured such that the rows correspond to the true classes, while the columns correspond to the predicted classes. The binary classification confusion matrix consists of components:

- True Positives (TP): whose true label is True and predicted as True.
- False Positives (FP): whose true label is False but predicted as True.
- True Negatives (TN): whose true label is False and predicted as False.
- False Negatives (FN): whose true label is True but predicted as False.

The models' performances on the new dataset are shown below by the confusion matrix.

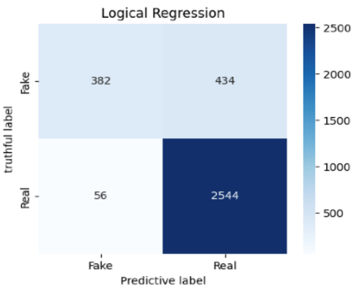


Figure 9: Logistic regression confusion metrics

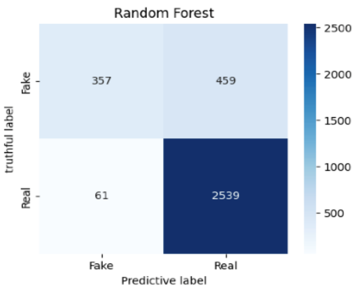


Figure 10: Random forest confusion metrics

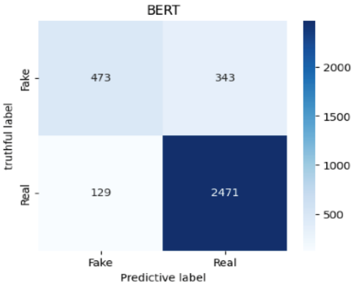


Figure 11: Bert fine-tuned confusion metrics

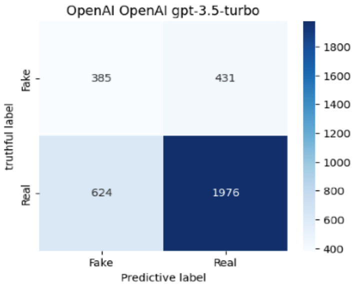


Figure 12: GPT confusion metrics

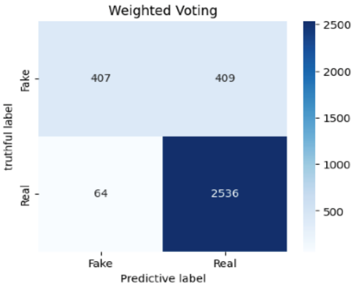


Figure 13: Weighted voting confusion metrics

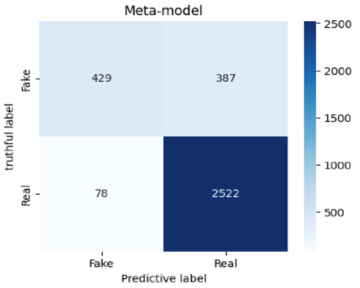


Figure 14: LR aggregation confusion metrics

4.5 Evaluation of Large Model Fine-Tuning

This experiment compared the classification performance of two models—**0.5B** and **1.5B**—at four training checkpoints (100, 500, 1000, and 1500). Evaluation metrics included **Accuracy**, **Precision**, **Recall**, and **F1-score**.

Overall Performance on Mainstream Samples

Overall, both model scales exhibited similar performance across the first three checkpoints (100–1000), achieving high accuracy and F1-scores on mainstream samples. For instance, at checkpoint 1000, the 0.5B model achieved an accuracy of 0.832 and an F1-score of 0.8945, while the 1.5B model recorded 0.8453 and 0.9022, respectively—indicating comparable results.

Impact of Model Size on Stability

At checkpoint 1500, the 0.5B model experienced a marked performance drop (accuracy falling to 0.4766, F1-score only 0.5671), suggesting possible overfitting or instability in the later training stages. In contrast, the 1.5B model maintained stability at this stage (accuracy 0.8681, F1-score 0.91), demonstrating stronger training robustness and sustained learning ability.

Observations on Metric Differences

From a metric perspective, the 0.5B model achieved its highest recall of 0.9362 around checkpoint 1000, with precision remaining in the 0.85–0.87 range. Meanwhile, the 1.5B model consistently achieved recall above 0.93, peaking at 0.97, and demonstrated a better balance between precision and recall.

Experimental Conclusions

- **Comparable accuracy**: On mainstream samples, both 0.5B and 1.5B models achieved similarly high accuracy and F1-scores.
- **1.5B is more robust**: The 1.5B model showed no performance degradation over longer training cycles, indicating good generalization ability.
- **0.5B is more prone to instability**: While it performed well in early stages, the 0.5B model was more susceptible to overfitting in later training due to limited capacity.
- **Recommendation**: If computational resources are limited and training is short-term, the 0.5B model can be a viable option. For deployment or long-term tasks, the 1.5B model is recommended to ensure stable performance.

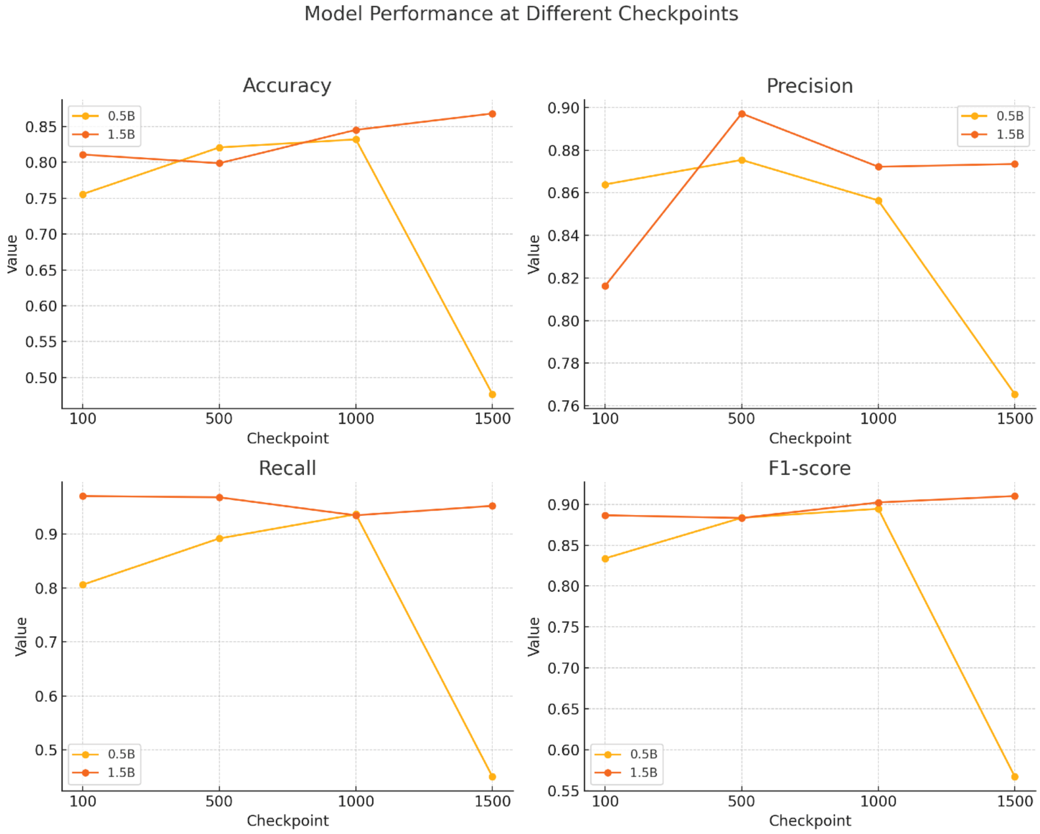


Figure 15 Checkpoint-wise Performance of Fine-Tuned Large Models

4.6 Comparative Analysis

Based on the experimental results, I compared the performance of the FIND ensemble learning framework with that of a fine-tuned large language model (LLM) in fake news detection.

The FIND ensemble framework integrates three structurally complementary classifiers, Random Forest, fine-tuned BERT, and a prompt-based large language model logistic regression or weighted voting. This approach effectively improves overall model stability and detection accuracy on mainstream samples, achieving performance comparable to that of a fine-tuned LLM in terms of overall accuracy. Its ensemble strategy is particularly suited for robust fusion in scenarios where individual model performances vary, making it well-suited for modular deployment in practical applications.

However, the overall performance of FIND shows certain limitations when handling long-tail cases. In the context of fake news detection, "long-tail news" refers to items with low dissemination volume and limited mainstream attention but large aggregate quantity. Such news items often have non-mainstream dissemination paths, niche audiences, and high verification difficulty. Compared to trending news, long-tail news frequently covers geographically remote regions, obscure disciplines, or topics with a sensational or curiosity-driven nature—making them more susceptible to exploitation by fake news creators seeking to bypass mainstream fact-checking mechanisms [46][47]. While ensemble strategies can partially mitigate errors from individual sub-models, they remain constrained by the capacity limits of base models and biases in the training data distribution, which may lead to instability in handling these complex cases.

In contrast, the fine-tuned LLM demonstrates stronger semantic generalization and reasoning capabilities in long-tail tasks. Through domain-specific instruction tuning, the LLM effectively learns the linguistic patterns and logical features of low-frequency samples. This enables it to better interpret ambiguous statements, identify sentiment tendencies, and detect factual inconsistencies. In our experiments, we observed that the fine-tuned LLM achieved a significantly larger accuracy improvement than the FIND model for low-frequency categories in the validation set, indicating its greater suitability for nuanced decision-making in complex text scenarios within fake news detection.

The figure below presents an example of detecting long-tail news.

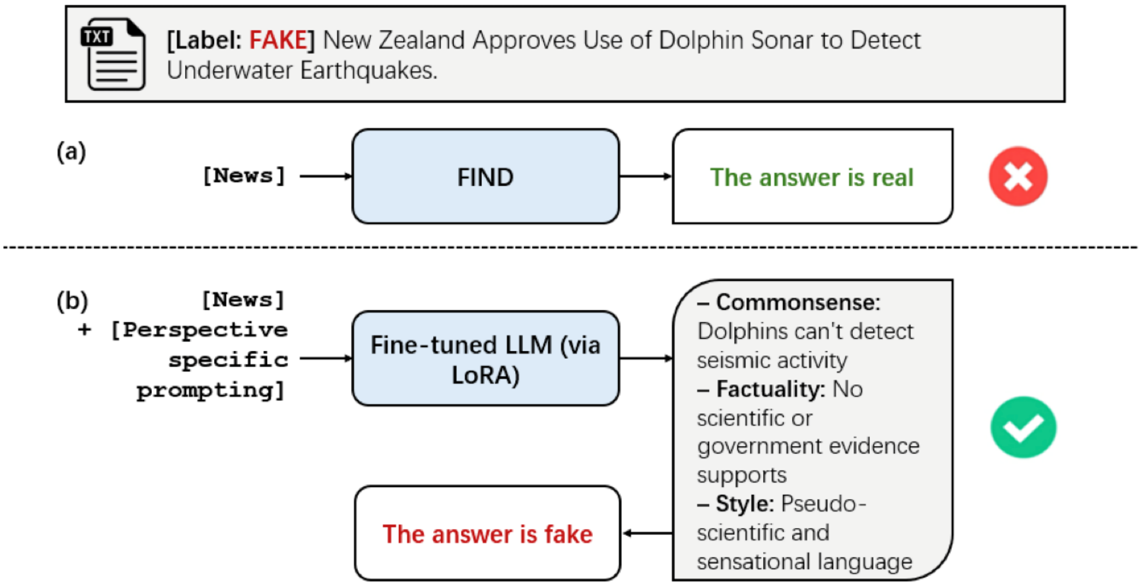

[Label: FAKE] New Zealand Approves Use of Dolphin Sonar to Detect Underwater Earthquakes.
(a)
[News]
FIND
The answer is real
(b)
[News] + [Perspective specific prompting]
Fine-tuned LLM (via LoRA)
– Commonsense: Dolphins can't detect seismic activity
– Factuality: No scientific or government evidence supports
– Style: Pseudo-scientific and sensational language
The answer is fake

Figure 16. Representative Examples of True and Fake News from the Dataset

The news headline *"New Zealand Approves Use of Dolphin Sonar to Detect Underwater Earthquakes"* represents a typical long-tail fake news case. On the surface, the story combines seemingly plausible elements such as technology, environmental protection, and earthquake monitoring, but in reality, a dolphin's sonar system is designed solely for short-range prey localization and lacks the capability to detect tectonic activity. Current scientific methods for underwater earthquake monitoring primarily rely on high-sensitivity instruments such as ocean-bottom seismometers and hydrophones, rather than biological sensing systems. Consequently, this news item not only contradicts basic scientific knowledge but, owing to its high technical specificity and non-mainstream dissemination channels, is also more difficult for the general public to identify as false. This underscores the unique challenges that long-tail news poses in the detection of misinformation.

V. FUTURE RESEARCH

We plan to integrate the proposed model into major social media platforms such as Twitter, Facebook, and WeChat. This will involve processing real-time data streams by leveraging platform APIs for information retrieval and live analysis. By implementing a real-time monitoring system, we aim to dynamically assess user-posted news content, thereby improving both the efficiency and accuracy of fake news detection. Furthermore, we will optimize the model for the specific characteristics of each platform to accommodate variations in user demographics and content types.

Future work will also focus on designing an interactive mechanism that enables users to label and provide feedback on the authenticity of news. This mechanism will not only supply the model with continuous training data but also allow it to self-improve by incorporating user knowledge and judgments. By integrating user feedback, the model can better adapt to the rapidly evolving social media landscape, thereby increasing its detection accuracy.

To further enhance detection performance, we plan to adopt a cross-modal learning approach that combines text, image, and audio data. By constructing a multimodal neural network, the model will be able to analyze the textual content of news alongside associated images and audio signals. For example, when processing video-based news, the model can extract key frames and background audio while jointly analyzing the accompanying text. This approach can help identify more complex patterns of misinformation, as fake news often exhibits cross-modal consistency.

REFERENCES

[1]. Guo, Z., Yu, K., Bashir, A. K., Zhang, D., Al-Otaibi, Y. D., & Guizani, M. (2022). Deep information fusion-driven POI scheduling for mobile social networks. IEEE Netw., 36 (4), 210–216. Retrieved from https://doi.org/10.1109/MNET.102.2100394

[2]. Allcott, H., & Gentzkow, M. (2017, May). Social media and fake news in the 2016 election. Journal of Economic Perspectives, 31 (2), 21136. Retrieved from https://www.aeaweb.org/articles?id=10.1257/jep.31.2.211 DOI: 10.1257/jep.31.2.211

[3]. Zhou, X., & Zafarani, R. (2020). A survey of fake news: Fundamental theories, detection methods, and opportunities. ACM Comput. Surv., 53 (5). Retrieved from https://doi.org/10.1145/3395046 DOI: 10.1145/3395046

[4]. News, B. (2016). China investigates search engine baidu after student's death. Retrieved 2024-10-31, from https://www.bbc.com/news/business-36189252

[5]. Coleman, A. (2020). 'hundreds dead' because of covid-19 misinformation. Retrieved 2024-10-31, from https://www.bbc.com/news/world-53755067

[6]. Asghar, M. Z., Habib, A., Habib, A., Khan, A., Ali, R., & Khattak, A. M. (2021). Exploring deep neural networks for rumor detection. J. Ambient Intell. Humaniz. Comput., 12 (4), 4315–4333. Retrieved from https://doi.org/10.1007/s12652-019-01527-4

[7]. Science., D. (n.d.). Dimensions ai. Retrieved from https://app.dimensions.ai (Accessed: 2024-10-28)

[8]. Brown, T. B. (2020). Language models are few-shot learners. arXiv preprint arXiv:2005.14165

[9]. Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2018). BERT: pre-training of deep bidirectional transformers for language understanding. CoRR, abs/1810.04805 . Retrieved from http://arxiv.org/abs/1810.04805

[10]. OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., . . . Zoph, B. (2024). Gpt-4 technical report. Retrieved from https://arxiv.org/abs/2303.08774

[11]. Hu, B., Sheng, Q., Cao, J., Shi, Y., Li, Y., Wang, D., & Qi, P. (2024). Bad actor, good advisor: Exploring the role of large language models in fake news detection. In Proceedings of the aaai conference on artificial intelligence (Vol. 38, pp. 22105–22113).

[12]. Chua, L. O. (1997). Cnn: A vision of complexity. International Journal of Bifurcation and Chaos, 7 (10), 2219–2425.

[13]. Sherstinsky, A. (2020). Fundamentals of recurrent neural network (rnn) and long short-term memory (lstm) network. Physica D: Nonlinear Phenomena, 404 , 132306

[14]. Qin, L., Chen, Q., Feng, X., Wu, Y., Zhang, Y., Li, Y., . . . Yu, P. S. (2024). Large language models meet nlp: A survey. arXiv preprint arXiv:2405.12819 .

[15] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., . . . Amodei, D. (2020). Language models are few-shot learners. CoRR, abs/2005.14165 . Retrieved from https://arxiv.org/abs/2005.14165

[16] Su, X., Cui, Y., Liu, A., Lin, X., Wang, Y., Liang, H., . . . Yu, Z. (2024). Daad: Dynamic analysis and adaptive discriminator for fake news detection. arXiv preprint arXiv:2408.10883 .

[17] Hu, B., Sheng, Q., Cao, J., Shi, Y., Li, Y., Wang, D., & Qi, P. (2024). Bad actor, good advisor: Exploring the role of large language models in fake news detection. In Proceedings of the aaai conference on artificial intelligence (Vol. 38, pp. 22105–22113).

[18]. Teo, T. W., Chua, H. N., Jasser, M. B., & Wong, R. T. (2024). Integrating large language models and machine learning for fake news detection. In 2024 20th ieee international colloquium on signal processing & its applications (cspa) (pp. 102–107).

[19]. Xue, J., Wang, Y., Tian, Y., Li, Y., Shi, L., & Wei, L. (2021). Detecting fake news by exploring the consistency of multimodal data. Information Processing & Management, 58 (5), 102610.

[20] Xu, R., & Li, G. (2024). A comparative study of offline models and online llms in fake news detection. arXiv preprint arXiv:2409.03067.

[21]. Pavlyshenko, B. M. (2023). Analysis of disinformation and fake news detection using fine-tuned large language model. arXiv preprint arXiv:2309.04704.

[22] Liu, Y., Zhu, J., Zhang, K., Tang, H., Zhang, Y., Liu, X., . . . Chen, E. (2024). Detect, investigate, judge and determine: A novel llm-based framework for few-shot fake news detection. arXiv preprint arXiv:2407.08952.

[23] as [17]

[24] as [21]

[25] as [17]

[26] Li, X., Zhang, Y., & Malthouse, E. C. (2024). Large language model agent for fake news detection. arXiv preprint arXiv:2405.01593.

[27] Wright, R. E. (1995). Logistic regression.

[28] Song, Y.-Y., & Ying, L. (2015). Decision tree methods: applications for classification and prediction. Shanghai archives of psychiatry, 27 (2), 130.

[29] Breiman, L. (2001). Random forests. Machine Learning, 45 (1), 5–32.

[30] as [29]
[31] Liaw, A., & Wiener, M. (2002). Classification and regression by randomforest. R News, 2 (3), 18–22.
[32] Cutler, D. R., Edwards Jr, T. C., Beard, K. H., Cutler, D. R., & Hess, K. T. (2007). Random forests for classification in ecology. Ecology, 88 (11), 2783–2792.
[33] Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2018). BERT: pre-training of deep bidirectional transformers for language understanding. CoRR, abs/1810.04805 . Retrieved from http://arxiv.org/abs/1810.04805
[34] Christian, H., Agus, M. P., & Suhartono, D. (2016). Single document automatic text summarization
using term frequency-inverse document frequency (tf-idf). ComTech: Computer, Mathematics and Engineering Applications, 7 (4), 285–294.
[35] Singh, A., Ehtesham, A., Kumar, S., & Khoei, T. T. (2024). Enhancing ai systems with agentic workflows patterns in large language model. In 2024 ieee world ai iot congress (aiiot) (pp. 527–532).
[36] Bristow, D. A., Tharayil, M., & Alleyne, A. G. (2006). A survey of iterative learning control. IEEE control systems magazine, 26 (3), 96–114.
[37] Amatriain, X. (2024). Measuring and mitigating hallucinations in large language models: amultifaceted approach.
[38] Li, X., Zhang, Y., & Malthouse, E. C. (2024). Large language model agent for fake news detection. ArXiv preprint arXiv:2405.01593 .
[39] Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., & Chen, W. (2021). Lora: Low-rank adaptation of large language models. CoRR, abs/2106.09685. Retrieved from https://arxiv.org/abs/2106.09685
[40] Yang, A., Yang, B., Hui, B., Zheng, B., Yu, B., Zhou, C., … Fan, Z. (2025). Qwen2.5 technical report. arXiv preprint arXiv:2412.15115. Retrieved from https://arxiv.org/abs/2412.15115
[41] as [39]
[42] Kaufmann, T., Weng, P., Bengs, V., & Hüllermeier, E. (2024). A survey of reinforcement learning from human feedback. Retrieved from https://arxiv.org/abs/2312.14925
[43] Plisson, J., Lavrac, N., Mladenic, D., et al. (2004). A rule based approach to word lemmatization. In Proceedings of is (Vol. 3, pp. 83–86).
[44] as [34]
[45] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., . . . Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of machine learning research, 21 (140), 1–67.
[45] Pavlyshenko, B. M. (2023). Analysis of disinformation and fake news detection using fine-tuned large language model. arXiv preprint arXiv:2309.04704
[46] Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake news detection on social media: A data mining perspective. ACM SIGKDD Explorations Newsletter, 19(1), 22–36. DOI: 10.1145/3137597.3137600
[47] Zhou, X., & Zafarani, R. (2020a). A survey of fake news: Fundamental theories, detection methods, and opportunities. ACM Comput. Surv., 53(5). Retrieved from https://doi.org/10.1145/3395046 DOI: 10.1145/3395046