Explore the Potential of LLMs in Misinformation Detection: An Empirical Study

Mengyang Chen[1,2], Lingwei Wei[1,*], Han Cao [1,2], Wei Zhou [1], Songlin Hu [1,2]

Institute of Information Engineering, Chinese Academy of Sciences [1]
School of Cyber Security, University of Chinese Academy of Sciences [2]
{chenmengyang, weilingwei, caohan, zhouwei, husonglin}@iie.ac.cn

Abstract

Large Language Models (LLMs) have garnered significant attention for their powerful ability in natural language understanding and reasoning. In this paper, we present a comprehensive empirical study to explore the performance of LLMs on misinformation detection tasks. This study stands as the pioneering investigation into the understanding capabilities of multiple LLMs regarding both content and propagation across social media platforms. Our empirical studies on eight misinformation detection datasets show that LLM-based detectors can achieve comparable performance in text-based misinformation detection but exhibit notably constrained capabilities in comprehending propagation structure compared to existing models in propagation-based misinformation detection. Our experiments further demonstrate that LLMs exhibit great potential to enhance existing misinformation detection models. These findings highlight the potential ability of LLMs to detect misinformation.

Introduction

Misinformation (Shu et al. 2017) [1] generally refers to false information that is spread deliberately to deceive people and cause severe negative impacts in the fields of national politics (Fisher, Cox, and Hermann 2016), economy (Vosoughi, Roy, and Aral 2018), and society (Faris et al. 2017). Existing misinformation detection methods mainly focus on news text and its propagation (engaged users, comments, retweet behaviors. etc). Text-based misinformation detection methods identify misinformation mainly by feeding extracted text features of the news into classifiers for classification (Castillo, Mendoza, and Poblete 2011; Ma et al. 2015, 2016; Yu et al. 2017; Luvembe et al. 2023; Hamed, Ab Aziz, and Yaakub 2023). Consequently, some researchers incorporate the propagation of the news such as users (Lu and Li 2020; Su et al. 2023) and propagation structure (Liu and Wu 2018; Bian et al. 2020; Wei et al. 2021; Wu and Hooi 2023; Chen et al. 2024) into the detection process, achieving promising performance.

Large language models (LLMs) possess rich world knowledge and human-like preferences (Bai et al. 2022), demonstrating comparable abilities to humans in many tasks (Nori et al. 2023; au2 and Katz 2022). In misinformation detection, previous studies (Hu et al. 2024; Huang and Sun 2023; Leite et al. 2023) have shown that ChatGPT (OpenAI 2022) exhibits slightly lower performance than small language models (SLMs) [2] when using text content for misinformation detection. However, they only focus on specific LLMs and overlook the assessment of LLMs' ability to leverage social context, which is particularly crucial for misinformation detection. Therefore, a more comprehensive evaluation of LLMs in exploiting content and propagation is still underexplored for the task.

In this paper, an empirical study is first conducted to assess the ability of LLMs to detect misinformation. We provide a more comprehensive analysis of LLMs on content and propagation-based misinformation detection. To explore the capabilities and potential of LLMs in misinformation detection, we divided the evaluation into two parts: an assessment of *LLM-based* detectors and an assessment of *LLM-enhanced* detectors.

In the evaluation of LLM-based detectors, we use different prompts to guide LLMs in misinformation detection directly. The results show that, under well-designed prompts, LLMs can achieve performance comparable to fine-tuned smaller models in text-based misinformation detection tasks. Different LLMs exhibit varying topic preferences and language preferences in tasks. In propagation-based misinformation detection, the performance of LLMs significantly lags behind existing detectors. This is because LLMs have difficulty understanding graph-structured propagation information through natural language alone.

In the evaluation of LLM-enhanced detectors, we utilize LLMs for data enhancement and feature enhancement, in combination with existing models for detection. The results indicate a significant improvement in detection performance. Specifically, LLM embeddings outperform the embeddings from mainstream text embedding models like BERT on most of the datasets, and their generative capabilities can simulate social user comments, enriching the propagation data

* Corresponding author.

[1] Both fake news and rumors both belong to misinformation. For convenience of reference, in this article, we use the term "news" to refer to information.

[2] SLMs refer to models with smaller parameter counts compared to LLMs (Hu et al. 2024)

available for detection.

The contributions of this work can be summarized as follows:

1) A comprehensive empirical study is conducted to evaluate LLMs in misinformation detection with diverse prompting methods. To the best of our knowledge, this is the first study to explore the understanding ability of multiple LLMs for both content and social context on social media.

2) We propose a novel evaluation framework that divides the analysis into LLM-based detectors and LLM-enhanced detectors, allowing for a deeper understanding of how LLMs perform independently and in combination with traditional models for misinformation detection.

3) Our study highlights LLMs' potential ability in misinformation detection, contributing to further exploration of LLMs' applicability in this task.

Related Work

Misinformation Detection

Misinformation detection aims to identify and judge the authenticity of news. Existing detection methods focus on the news text and its propagation on social media.

Text-based methods involve extracting features from the news for classification. Early works rely on crafted features (Castillo, Mendoza, and Poblete 2011; Ma et al. 2015; Holan 2016), which required significant manual effort and domain expertise. With the development of deep learning, many works have employed neural networks to automatically learn high-level features (Ma et al. 2016; Ma, Gao, and Wong 2019; Cheng, Nazarian, and Bogdan 2020; Yu et al. 2017; Vaibhav, Annasamy, and Hovy 2019). Recent works employ pre-trained language models (Devlin et al. 2019; He et al. 2020; Hu et al. 2022) to learn a better representation from content for detection (Kaliyar, Goswami, and Narang 2021; Jwa et al. 2019; Hamed, Ab Aziz, and Yaakub 2023; Luvembe et al. 2023).

propagation-based methods introduce user information (Dong et al. 2018; Nguyen et al. 2020; Yuan et al. 2020; Dou et al. 2021) and propagation structure (Ma, Gao, and Wong 2017, 2018; Bian et al. 2020; Song, Shu, and Wu 2021; Hu et al. 2021a; Wei et al. 2022a; Wu and Hooi 2023) to determine authenticity. To solve the issues raised by incomplete and noise propagation, some works use the contrastive learning (Ma et al. 2022; He et al. 2021; Sun et al. 2022), learned uncertainty (Wei et al. 2021, 2022c; Chen et al. 2024), and propagation reconstruction (Wei et al. 2022b; Wu and Hooi 2023) for robust structural representations.

Large Language Models

Large language models (LLMs) refer to language models that are trained on large-scale text corpora and contain hundreds of billions (or more) parameters, such as GPT-3, PaLM, LLaMA, etc (Zhao et al. 2023). Researchers have found that "emerging capabilities" enable LLMs to significantly perform well while demonstrating special abilities that smaller models lack, such as contextual learning, instruction following, and step-by-step reasoning (Fan et al. 2023), showing broad application prospects in many fields such as medicine, education, and finance (Naveed et al. 2023).

LLM has received increasing attention in misinformation detection (Chen and Shu 2024). Some works evaluate LLMs in using text content for misinformation detection (Hu et al. 2024; Huang and Sun 2023; Choi et al. 2023). They find that LLMs are not as good as existing text-based detection models and highlight the supporting role of LLMs in detecting misinformation, including professional analyses for news (Hu et al. 2024; Huang and Sun 2023), different writing styles of LLMs (Wu, Guo, and Hooi 2024) and propagation generation for misinformation(Nan et al. 2024; Wan et al. 2024). Despite these advancements, the performance of multiple LLMs and their abilities to utilize propagation to detect misinformation remain unexplored.

Preminarlies

In this section, we present concepts, notations, and problem settings used in the work.

Misinformation Detection is to verify the authenticity of a given news article, we take misinformation detection as a binary classification problem, where each sample is annotated with a ground truth label indicating its authenticity. In text-based misinformation detection, each sample consists of a single news article so it can be seen as text classification. In propagation-based misinformation detection, each sample includes not only the news content but also user comments and the propagation structure formed between the news and the comments.

Formally, in propagation-based misinformation detection, Dataset $\mathcal{D}$ consists of N samples and each sample is represented by $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V} = \{r, c_1, ..., c_N\}$ represents the features of the news r and its comments $(c_1, ..., c_N)$, $\mathcal{E}$ represents a set of explicit interactive behaviors, e.g., retweet. The task objective of misinformation detection is to classify samples and determine whether the news is true (labeled as 1) or false (labeled as 0) by a binary classifier.

LLM-based Detectors As shown in Figure 1 (a), LLM-based detectors involve the direct utilization of large language models (LLMs) for misinformation detection. Specifically, we provide task-specific prompts to guide the LLMs in performing the detection task.

LLM-enhanced Detectors LLM-enhanced detectors, on the other hand, use existing detection models as the core, with LLMs serving for data enhancement. As shown in Figure 1 1 (b). This is done in two forms: 1) Using LLMs as the backbone for text embeddings, SLMs utilize the embeddings generated by LLMs for detection. 2) Using synthetic data generated by LLMs as a supplement to existing data for SLMs.

Evaluation of LLM-based Detectors

In this section, we evaluate the performance of LLMs as detectors in misinformation detection. Specifically, we guide LLMs towards detecting misinformation through various prompting methods and Instruction-Tuning.

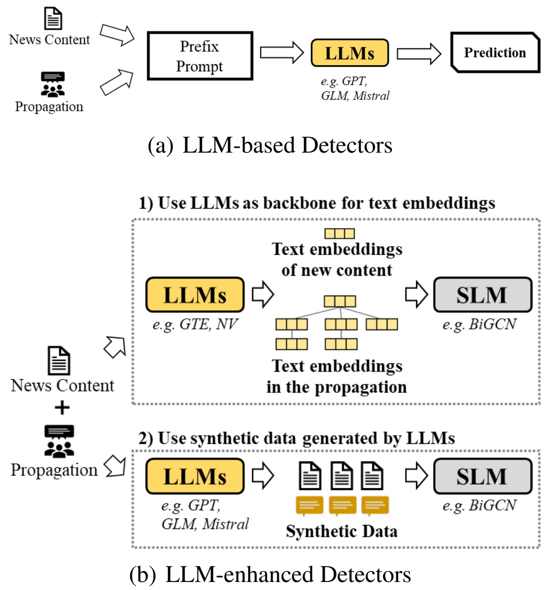


Figure 1: Two paradigms for utilizing LLMs in misinformation detection.

Evaluation Settings

Datasets We experiment on eight representative misinformation detection benchmark datasets, consisting of three datasets containing news content and five datasets containing content and propagation. These datasets primarily come from news media and social platforms, covering topics such as politics, military, entertainment, society, and health. The statistics of datasets are listed in Table 7 . **FakeNewsNet** (Shu et al. 2018) is a dataset compiled from GossipCop and PolitiFact, labeled as true or false by experts, primarily covering political and entertainment news from 2015 to 2017. **LTCR** (Ma et al. 2023) is a Chinese rumor detection dataset with long sentences, sourced from official media during the COVID-19 pandemic. **FakeNewsDataset23** (FND23) is collected from fact-checking platforms such as Snopes and PolitiFact. It contains English news titles that were published after 2019, which are labeled as true or false. **Twitter**, created by Siska et al. (2024), contains tweets published on Twitter[3], and each tweet is annotated with true of false. **Twitter covid** (Siska et al. 2024) is a diverse COVID-19 healthcare misinformation dataset, including content and user social engagement. **PHEME5** (Zubiaga, Liakata, and Procter 2016) contains collections of rumors and non-rumors released on Twitter during 5 emergency events between 2014 and 2016. **CED** (Song et al. 2018) contains Chinese rumor data scraped from Weibo, including forwarding and comment information related to the original Weibo posts, which are divided into rumors and non-rumors. **Weibo covid** (Siska et al. 2024) contains rumors about COVID-19 scraped from Weibo, which are divided into rumors and non-rumors.

Following the dataset split, we divide the datasets into training, validation, and test sets according to a ratio of 7:1:2.

[3]In July 2023, Twitter has been rebranded to X.

Comparison Methods

Selection of SLMs For text-based detection models, following (Hu et al. 2024), we selected 2 representative small models for text-based and propagation-based methods and 5 LLMs to evaluate LLM performance. we choose **Multi-Layer Perceptron (MLP)** and **EANN** (Wang et al. 2018) fine-tuned on the training set as comparison methods. For propagation-based detection models, we take Random Selection, Graph Convolutional Network (GCN), and a representative propagation-based detection method (i.e., **Bi-GCN** (Bian et al. 2020)) as baseline methods for comparison.

Selection of LLMs For LLMs, we experiment with GPT-3.5-turbo, GLM, Mistral, Qwen and Vicuna for comparison. **GPT-3.5-turbo** [4] is the most capable and cost-effective model in the GPT-3.5 family and in this experiment we use the default version *gpt-3.5-turbo-0125*. **GLM** (Du et al. 2021) is a general language model pre-trained with an autoregressive blank-filling objective and has been optimized for Chinese question-answering and dialogue. In the experiments, we use *ChatGLM3-6B* [5] . **Qwen** (Yang et al. 2024) is a language model series including decoder language models of different model sizes developed by Alibaba Cloud. In the experiments, we use *Qwen1.5-7B-Chat* [6] . **Mistral** (Jiang et al. 2023)is the first dense model released by Mistral AI, designed to be ideal for experimentation, customization, and rapid iteration. We utilized *Mistral-7B-Instruct-v0.2* [7] for the evaluation. **Vicuna** (Zheng et al. 2023) is a chat assistant trained by fine-tuning Llama 2 on user-shared conversations collected from ShareGPT. The version we use in the evaluation is *vicuna-7b-v1.5* [8] .

Prompt Settings In this paragraph, we will clarify the prompt settings we use in evaluating the performance of LLMs as detectors in misinformation detection. As shown in Table 8 and Table 9 , to explore the potential of LLMs, and consider the critical role that prompt settings play in their task performance, we designed three **Vanilla prompts** (Shown in Table 8). Based on these Vanilla prompts, we developed additional prompts such as Task Prompt and Refine Prompt[9].

For text-based detection, we utilize the following prompt settings: **Task Prompt** (Yin et al. 2023) involve incorporating task-related definitions into instructions to enhance the performance of LLMs on specific tasks. We introduce the definition of misinformation in vanilla prompts. **Chain-of-Thought** (Wei et al. 2023) (CoT) prompt refers to a series of logically related thinking steps or ideas that are connected, forming a complete thinking process. We guide LLMs to carefully think through the detection process and then provide a judgment by adding "think step by step" in each prompt.

[4]For detailed information about this model, please refer to https://platform.openai.com/docs/models/gpt-3-5

[5]https://huggingface.co/THUDM/chatglm3-6b

[6]https://huggingface.co/Qwen/Qwen1.5-7B-Chat

[7]https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.2

[8]https://huggingface.co/lmsys/vicuna-7b-v1.5

[9]Each type of these prompts also has three variants developed based on the Vanilla prompts, and we only present one variant of each type in Table 8 and Table 9 .

Methods	FND23		FNN		LTCR	
	Accuracy	Macro-F1	Accuracy	Macro-F1	Accuracy	Macro-F1
BERT	85.85	55.79	81.42	<u>71.34</u>	**95.63**	<u>93.98</u>
EANN	**89.23**	47.34	<u>83.24</u>	**75.45**	<u>95.38</u>	**94.22**
GPT3.5-Turbo						
Vanilla prompt	82.56	55.62	58.87	56.06	82.60	77.20
Task Prompt	87.50	52.76	60.18	57.17	83.77	73.95
COT	81.10	58.84	73.98	57.98	81.47	65.07
RA Prompt	67.08	51.74	68.29	59.05	80.64	73.99
GLM						
Vanilla prompt	85.76	54.58	64.76	56.9	86.87	79.59
Task Prompt	88.37	46.91	58.72	53.84	88.62	82.86
COT	87.79	46.75	69.41	60.04	86.43	78.19
RA Prompt	85.47	55.38	50.09	48.06	80.09	76.13
Qwen						
Vanilla prompt	76.16	55.63	65.14	59.90	90.81	85.83
Task Prompt	78.78	58.32	67.17	62.15	90.81	85.83
COT	78.20	56.48	82.82	49.28	90.37	85.49
RA Prompt	62.21	50.00	67.77	61.22	90.37	84.55
Mistral						
Vanilla prompt	80.56	**61.74**	62.12	60.70	68.93	48.57
Task Prompt	83.24	<u>60.79</u>	81.50	62.03	75.68	47.97
COT	86.81	53.57	70.29	63.03	77.20	52.32
RA Prompt	83.33	55.87	**83.60**	50.41	87.72	76.05
Vicuna						
Vanilla prompt	83.07	49.41	70.42	50.68	75.17	49.72
Task Prompt	85.93	50.19	73.65	49.29	76.09	51.09
COT	<u>89.05</u>	47.10	74.92	44.04	75.56	49.29
RA Prompt	82.87	50.11	71.01	51.22	75.52	53.21

Table 1: Results on text-based misinformation detection. The best results for each dataset are highlighted in **bold**, and the second-best results are <u>underlined</u>.

Reason-aware Prompt (Huang et al. 2024) helps LLMs assess misinformation by providing the underlying causes of misinformation within the prompt. Moreover, we employ Few-shot learning for LLMs, where the number of shots $N = \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$, evaluating its ability to generalize and adapt to detect misinformation. In N-shot learning ($N > 1$), the model is provided with N-shot exemplars from the training set, to aid its adaptation to the task.

For propagation-based detection, we utilize the following prompt settings: **Refine Prompt**: We enhance the vanilla prompts by instructing the LLMs to filter out comments unrelated to the news in the propagation information, reducing the impact of redundant information on the model. **Format-Graph-Input (FGI)**: Inspired by the work of (Guo et al. 2023), we describe the news propagation graph using formal graph terminology (e.g., nodes, edges) and then incorporate this into the vanilla prompts.

Instruction-Tuning For the Instruction-Tuning, we construct instruction-tuning datasets from the original datasets to fine-tune **LLaMa2-7B** (Touvron et al. 2023), an open-source LLM developed by Meta AI with an upgraded size and performance for various NLP tasks. In the experiments, the temperature and top-p ratio are set to 0.6 and 0.9. We fine-tune LLaMa2-7B by using the widely adopted and efficient Low-Rank Adaptation (LoRA) technique (Hu et al. 2021b).

Evaluation Metrics. We run each prompt three times and report the average results. We use accuracy and macro-average F1 score as the evaluation metrics for the evaluation. For the evaluation of each prompt of LLMs, we select the best result from its three variants.

For the details of the implementation, please refer to the Appendix.

Overall Performance on Misinformation Detection

We evaluate the overall performance of text-based misinformation detection on FND23, FakeNewsNet, and LTCR, and that of propagation-based misinformation detection on the five datasets. The accuracy scores are presented in Table 1 and Table 2 . From the results, we can observe that:

The performance of LLMs in text-based misinformation detection is impressive. On the FND23 and FakeNewsNet datasets, LLMs achieve performance comparable to smaller models with specific prompts. For instance, on the FND23 dataset, Vicuna with the COT prompt achieves near-optimal accuracy, while Mistral with the RA prompt achieves optimal accuracy on the FNN dataset. This suggests that large models can leverage their extensive training data to make accurate judgments in misinformation detection. However, LLMs still lag behind SLMs on the LTCR dataset. We also found that although LLMs achieve a higher accuracy rate, the macro average F1 score remains low. This is due to the imbalance of positive and negative instances in the data (refer to Table 7), whereas SLMs eliminate this impact through training.

Incorporating task-related information enhances the detection performance of LLMs. From Table 1 , we can observe that when LLMs use task-specific prompts, their detection accuracy improves compared to vanilla prompts. This suggests that including a definition of misinformation in the prompt can help LLMs better understand the task.

Chain-of-Thought prompt and RA Prompt does not unequivocally enhance the performance of LLMs on misinformation detection. Some LLMs get reduced performance with the CoT Prompt compared to the Vanilla Prompt, showing difficulty for LLMs in reasoning tasks without extra cues. Like the COT Prompt, the RA Prompt failed to improve the performance of LLMs across all datasets equally. This reflects the limitations of current language models when dealing with multi-step reasoning tasks.

LLMs possess certain but limited understanding abilities for propagation. In Table 2 2, the detection performance of LLMs is better than Random. However, the performance of prompt-tuning LLMs is significantly inferior to that of the small model (i.e., Bi-GCN). When only news text is provided, LLMs perform poorly in detection, indicating that additional information is needed for better detection. Although including comments and propagation structure can improve detection performance, the enhancement is often limited. Specifically, we find that when using a vanilla prompt, providing comment content to LLMs generally yields better results than additionally supplying propagation structure in most cases. Moreover, using different prompts to describe the propagation structure can significantly impact the performance of LLMs.

The design of prompts has a significant impact on LLM's performance in detection. On two benchmarks, we find that different prompts have a substantial impact on LLM performance. For instance, on the FakeNewsNet dataset, using the RA prompt increases Mistral's detection accuracy by 21.48% compared to using Vanilla Prompts. Similarly, in propagation-based benchmarks, prompts significantly affect how LLMs understand propagated information. For exam-

Methods	Inputs			Twitter		Twitter covid		CED		Weibo covid		PHEME5	
	News	Comments	Structure	Accuracy	Macro-F1	Accuracy	Macro-F1	Accuracy	Macro-F1	Accuracy	Macro-F1	Accuracy	Macro-F1
Random	✓	✓	✓	50.00	50.00	50.00	50.00	50.00	50.00	50.00	50.00	50.00	50
GCN†	✓	✓	✓	77.20	76.61	76.67	75.08	83.04	81.92	91.97	90.47	80.31	75.80
Bi-GCN†	✓	✓	✓	82.34	79.58	77.50	74.42	93.45	93.22	91.57	90.30	81.48	79.71
GPT-3.5-turbo	✓	×	×	52.17	52.17	41.25	40.12	48.28	48.28	43.37	41.19	54.80	51.59
	✓	✓	×	57.33	55.50	47.62	45.65	65.36	65.26	37.18	35.90	46.33	46.10
Vanilla prompt	✓	✓	✓	51.29	51.18	56.36	**56.00**	57.42	53.96	48.10	45.88	**66.32**	48.44
Refine	✓	✓	✓	56.58	56.39	**58.18**	54.32	60.60	58.63	**49.37**	**49.16**	63.49	49.28
FGI	✓	✓	✓	56.71	**56.64**	56.64	53.75	**74.77**	**74.66**	44.58	44.38	60.96	**57.61**
GLM	✓	×	×	43.97	42.74	56.25	55.06	57.23	55.49	60.24	59.08	58.66	52.73
	✓	✓	×	47.53	44.00	**61.19**	**59.15**	**70.80**	**70.79**	**68.29**	**68.12**	60.17	**55.98**
Vanilla prompt	✓	✓	✓	50.25	46.15	54.72	51.82	57.53	56.3	64.10	62.53	**68.52**	47.16
Refine	✓	✓	✓	52.43	46.19	60.00	52.38	56.13	55.95	49.37	49.29	60.44	53.45
FGI	✓	✓	✓	**57.41**	**57.04**	48.57	46.43	52.09	49.38	57.69	50.91	56.91	50.25
Qwen	✓	×	×	59.05	49.94	57.50	43.33	48.08	37.22	61.45	40.82	32.47	25.93
	✓	✓	×	61.95	**55.15**	55.56	**45.91**	**80.10**	**80.05**	**68.42**	**56.32**	47.72	**47.71**
Vanilla prompt	✓	✓	✓	57.01	37.29	52.63	34.48	71.75	67.32	66.22	39.84	63.91	41.36
Refine	✓	✓	✓	54.81	36.35	55.26	40.35	66.67	65.22	64.86	39.34	60.29	42.49
FGI	✓	✓	✓	56.44	37.88	**58.75**	37.01	56.19	36.67	66.27	39.86	**67.70**	41.13
Mistral	✓	×	×	56.00	54.56	50	49.97	62.12	59.97	50.82	48.48	33.74	28.64
	✓	✓	×	57.22	55.27	41.38	29.27	65.03	64.36	50.00	49.43	**53.51**	**52.77**
Vanilla prompt	✓	✓	✓	48.84	46.76	**61.54**	**60.61**	**76.25**	**74.35**	41.27	40.73	52.14	52.04
Refine	✓	✓	✓	47.31	43.56	45.45	41.07	66.60	60.22	33.87	30.83	51.96	51.66
FGI	✓	✓	✓	**60.75**	**60.47**	44.64	44.63	74.77	74.66	**54.55**	**54.49**	41.31	38.00
Vicuna	✓	×	×	53.68	38.06	48.75	48.74	53.86	53.65	56.63	**52.30**	**53.71**	**51.43**
	✓	✓	×	54.46	**52.08**	48.00	47.67	56.20	**55.28**	**58.11**	50.51	48.18	47.45
Vanilla prompt	✓	✓	✓	**55.33**	47.76	41.33	40.04	48.48	44.35	40.85	38.08	47.26	46.57
Refine	✓	✓	✓	55.09	51.73	39.29	38.02	**56.44**	54.84	53.62	49.82	45.03	44.82
FGI	✓	✓	✓	49.51	49.50	**51.67**	**51.00**	46.03	46.02	54.24	51.51	51.13	48.62

Table 2: Results on propagation-based misinformation detection. The best result for each method on each dataset is highlighted in **bold**. The **Structure** in Inputs means propagation structure of news.

ple, on the CED dataset, the accuracy of Mistral in detecting propagation-based misinformation differs by 9.65% between using the Vanilla prompt and the Refine prompt. Furthermore, the way propagation information about the news is described (Vanilla and FGI) also affects how LLMs understand the propagated information. This indicates that careful design of prompts is essential when using LLMs for misinformation detection.

The performance of LLMs varies across different languages. We find that Chinese LLMs (GLM and Qwen) perform better overall than the other three LLMs on Chinese datasets (LTCR, CED, and Weibo COVID). This may be because GLM and Qwen are specifically optimized for the Chinese language and context during their training, resulting in higher accuracy in handling Chinese misinformation.

Further Analysis

In this section, we explore the effects of Few-shot Learning and Instruction-Tuning on LLMs in text-based misinformation detection.

Then, through case analysis, we attempt to summarize the factors influencing LLMs in misinformation detection.

Few-shot learning does not consistently enhance the performance of LLMs. As shown in Figure2, the performance of most LLMs (i.e., GPT-3.5-turbo and GLM) tends to decline with the initial few examples, it then gradually improves as the number of examples increases. Surprisingly, compared to the zero-shot setting, **most LLMs under one-shot learning show poor detection performance**. It indicates that when the number of training instances is insufficient, Few-shot learning fails to make the LLM fully understand the misinformation detection task.

Instruction-tuning could help LLMs in misinformation detection. The comparison of LLaMa2-7B's performance on eight datasets before and after instruction-tuning is shown in Table 3and Table4. The results show that LLaMa2-7B gains significant performance improvements after instruction-tuning. However, the performance of the fine-tuned LLaMa is still limited compared to SLM, especially in propagation-based misinformation detection. Moreover, the fine-tuned LLaMa2-7B performs poorly on Twitter, Twitter covid, and Weibo covid with small data sizes, which may suggest challenges for instruction-tuning in low-resource scenarios.

To explore the relationship between the volume of fine-tuning data and the performance (accuracy) of LLaMA2-7B after instruction-tuning, we conducted experiments on all datasets except Twitter, Twitter covid, and Weibo covid due to their limited data volumes. Figure 3 illustrates that the performance of the fine-tuned LLaMA2-7B improves with an increase in training data volume, whereas insufficient data leads to suboptimal results.

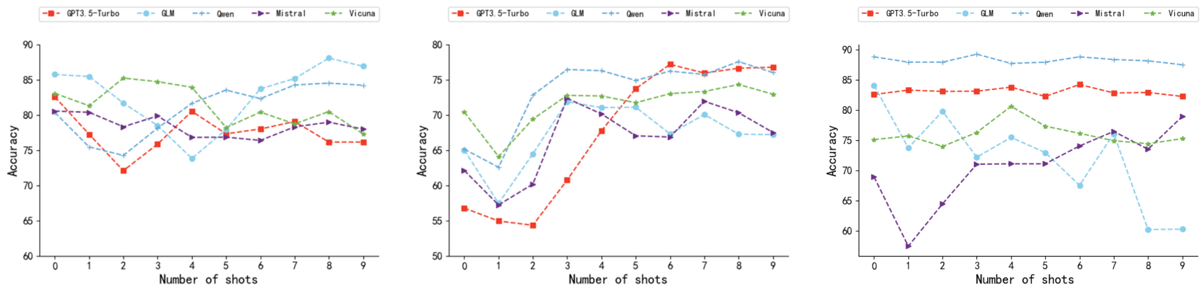
Models	FND23	FakeNewsNet	LTCR
MLP†	85.85	81.42	95.63
EANN†	89.23	83.24	95.38
LLaMa2-7B	54.36	67.69	60.26
LLaMa2-7B†	89.83	84.35	89.08

Table 3: Accuracy scores (%) of LLaMa2-7B on text-based misinformation detection. † indicates the model is fine-tuned on the full training set.

What affects LLMs in misinformation detection.

Through analysis, We have currently identified four factors that may affect the detection capabilities of LLMs:

1) The limitation of input length: The input content to be detected exceeds the input length limitations of LLMs, resulting in the LLMs only being able to process a limited portion of the input. In our experiments, we find that LLMs often encounter errors during detection due to the length of



(a) FND23 (b) FakeNewsNet (c) LTCR

Figure 2: Few-shot Learning results on three text-based misinformation detection datasets.

Methods	Inputs News	Inputs Comments	Inputs Relations	Twitter	Twitter coivd	CED	Weibo covid	PHEME5
Random	-	-	-	50.00	50.00	50.00	50.00	50.00
GCN†	✓	✓	✓	77.20	76.67	83.04	91.97	80.31
Bi-GCN†	✓	✓	✓	82.34	77.75	93.45	91.57	81.48
LLaMa2-7B	✓	×	×	43.54	36.54	51.18	37.18	48.34
	✓	✓	×	39.65	34.10	58.16	39.26	44.10
	✓	✓	✓	34.58	36.65	54.23	35.32	43.66
LLaMa2-7B†	✓	×	×	71.22	66.55	85.45	66.22	83.20
	✓	✓	×	65.41	61.54	83.58	64.86	86.19
	✓	✓	✓	67.86	58.24	89.43	67.45	87.77

Table 4: Accuracy scores (%) of LLaMa2-7B on propagation-based misinformation detection. We tune LLMs with vanilla prompts. † indicates the model is fine-tuned on the full training set.

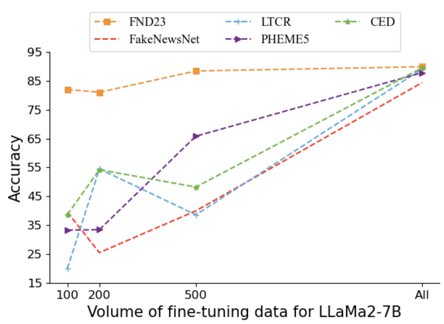


Figure 3: Accuracy scores (%) of instruction-tuning LLaMa2-7B against different training data size.

the input tokens exceeding the model's maximum context length. As shown in Table 7, the average token numbers of the Twitter covid dataset is approximately 78885, which has exceeded the maximum context length of Vicuna (4096 tokens). So we have to truncate the inputs to comply with the maximum context length of the LLMs;

2) Hallucination: The hallucinations of LLMs may affect their analysis of the content to be detected and lead them to make wrong judgments. We guide LLMs to generate judgments and analyses on the authenticity of the news by prompts. We then manually inspected the erroneous judgments and analyses, summarizing the following three reasons, as shown in Table 11. According to Ji et al. (2023).'s definition, these three reasons can be categorized as illusions or misconceptions of LLMs. This is consistent with the conclusions from the study by Hu et al. (2024);

3) Variability in the authenticity of news: The changing authenticity of news over time may confuse LLMs. The accuracy of all LLMs decreases on the FakeNewsNet dataset compared with that on the FND23 dataset. We find that this discrepancy may be caused by the timeliness of the news through analysis of the samples misjudged by LLMs. As shown in Figure 4, the authenticity of some news has changed over time, which contributes to the decreased accuracy of LLMs in detecting misinformation. Moreover, SLMs essentially learn the mapping from text to labels during training, so the changes in the truthfulness of the news do not significantly impact their judgment;

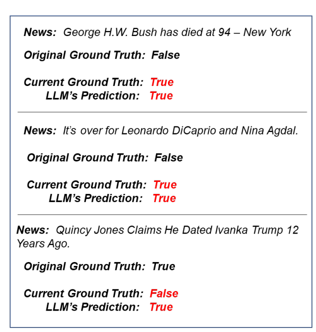


Figure 4: Changes in news authenticity lead to LLMs' misjudgment (Taking GPT-3.5-turbo as an example).

4) Topic bias: We use *GPT3.5-Turbo-0125* to perform topic classification on news from FND23, FNN, and LTCR, categorizing them mainly into a) political and military, b) technology and education, c) economics and finance, d) society, e) environment and health, f) entertainment. We explored the accuracy of LLMs in detecting misinformation across different categories of news, and the results are shown in Figure 5. It can be observed that LLMs also exhibit certain topic biases, they are more adept at judging news related to environmental and political topics compared to entertainment and social news.

Dicussion

This section primarily evaluates the ability of LLMs to act as predictors for misinformation detection through prompt-based experiments. The results show that in text-based misinformation detection, LLMs can perform comparably to or even surpass SLMs when given specific prompts. Including a definition of misinformation within the prompts helps LLMs better understand the task, thereby improving their performance. However, the prompt-based learning approach is not suitable for propagation-based misinformation detection. This is because LLMs struggle to comprehend propagation information described in natural language, and propagation data, being a topological structure, is not well-suited to be input into the model as a sequence. We also find that LLMs are not well-suited for detecting outdated news, as the truthfulness of such news may change over time, affecting the LLMs' label predictions. Additionally, prompting LLMs to reason step-by-step or providing key indicators of misinformation does not always enhance their performance. This is likely due to the inherent hallucinations within LLMs, which hinder their ability to correctly choose and summarize reasoning paths.

Evaluation of LLM-enhanced Detectors

In this section, we explore the potential of LLMs to enhance existing misinformation detection models on the previous tasks(text-based and propagation-based misinformation detection). We primarily consider utilizing LLMs for data enhancement on misinformation datasets to improve the judgment capabilities of existing detection models. Specifically, there are two approaches for LLMs to enhance news data. One is the feature-level enhancement, where LLMs leverage their extensive corpus knowledge to embed and enhance text features. The other approach is text-level enhancement, where LLMs use their generative abilities to produce new text, such as news analysis or commentary.

Evaluation Settings

Comparision Methods For feature enhancement using LLMs, we compare gte-Qwen2-1.5B-instruct as the encoder with BERT. gte-Qwen2-1.5B-instruct is a recently released General Text Embedding model designed to generate versatile and high-quality embeddings for various text-based tasks. We choose the SLMs in Section to evaluate the features generated by LLMs. Specifically, **w/ BERT** indicates that the text embeddings used by the model come from BERT. **w/NV** and **w/Qwen** indicates that the text embeddings come from *NV-Embed-v2* [10] and *gte-Qwen2-1.5B-instruct* [11] . For text-level enhancement, we select ARG, GenFEND, and DELL as the comparison methods. **ARG** is a framework that combines large language models (LLMs) with small language models (SLMs), it integrates the reasoning information from LLMs to assist SLMs in making more accurate judgments in news analysis. **GenFEND** is a generative feedback-enhanced misinformation detection framework. It provides LLMs with diverse user profiles to generate comments from different groups and aggregates these generated comments for improved detection. We use BERT as the backbone of GenFEND. **DELL** is a misinformation detection framework that integrates LLMs through a process involving generating news reactions to simulate diverse user perspectives, producing explanations to enrich context with sentiment and stance, predicting news with synthetic reactions using different expert Proxy Tasks, and merging expert predictions for a comprehensive judgment.

Methods	FND23		FNN		LTCR	
	Accuracy	Macro-F1	Accuracy	Macro-F1	Accuracy	Macro-F1
MLP						
w/BERT	85.85	55.79	81.42	71.34	95.63	93.98
w/DeBERTa	85.90	52.76	82.24	72.50	95.63	93.98
w/NV	87.60	66.01	85.04	79.05	94.32	82.05
w/Qwen	86.05	62.75	84.19	77.28	**100.00**	**100.00**
EANN						
w/BERT	89.83	47.32	83.04	82.14	95.63	93.562
w/DeBERTa	89.83	47.32	84.24	83.84	94.24	93.86
w/NV	89.89	47.32	**89.24**	**85.54**	93.89	91.32
w/Qwen	89.93	47.32	88.56	80.04	**100.00**	**100.00**
ARG	**90.12**	50.17	85.00	77.52	96.36	95.12
ARG-D	89.83	**62.94**	84.35	79.36	96.36	95.12

Table 5: Results on text-based misinformation detection. The best results for each dataset are highlighted in **bold**.

Implementation Details To eliminate the impact of language differences for a fair comparison, we use the *bert-base-uncased* version of BERT for **w/ BERT** methods and the three text-level enhancement methods (ARG, GenFEND, and DELL) to get textual features on English datasets, and the *bert-base-chinese* version to get textual features Chinese datasets[12]. To get the textual features from *NV-Embed-v2* and *gte-Qwen2-1.5B-instruct*, we set the task prompt as "*Given a sentence, encode it into an embedding*." For more details on the implementation of SLMs and the three text-level enhancement methods, please refer to Appendix . We use accuracy and macro-average F1 score as the evaluation metrics for each model.

Overall Performance of LLM-enhanced Detectors

We evaluated the performance of LLMs as augmenters in previous tasks, and the results are shown in Table 5 and Table 6. From the results, we have the following observations:

1) **In most cases, feature enhancement based on LLMs can improve the detection performance of existing mod-**

[10]https://huggingface.co/nvidia/NV-Embed-v2

[11]https://huggingface.co/Alibaba-NLP/gte-Qwen2-1.5B-instruct

[12]We also use the *deberta-base* and *deberta-base-chinese* for **w/DeBERTa** methods on English and Chinese dataset separately.

Methods	Twitter		Twitter covid		CED		Weibo covid		PHEME5	
	Accuracy	Macro-F1	Accuracy	Macro-F1	Accuracy	Macro-F1	Accuracy	Macro-F1	Accuracy	Macro-F1
GCN										
w/BERT	77.20	76.61	76.67	75.08	83.04	81.92	91.97	90.47	80.31	75.80
w/DeBERTa	76.19	75.74	75.17	75.65	85.77	83.95	90.32	89.78	82.02	79.14
w/NV	78.87	75.84	77.92	75.58	54.57	34.51	91.17	88.42	81.48	78.67
w/Qwen	78.87	75.84	77.50	74.42	62.98	62.15	**96.79**	**95.37**	82.00	79.51
BiGCN										
w/BERT	82.34	79.58	77.50	74.42	**93.45**	**93.22**	91.57	90.30	81.48	79.71
w/DeBERTa	82.37	80.91	78.72	77.84	93.45	93.22	92.68	91.88	79.65	79.56
w/NV	**89.18**	**88.98**	76.25	75.38	54.82	43.50	91.578	90.07	**85.27**	**82.08**
w/Qwen	85.79	81.75	75.00	74.42	61.21	61.12	93.98	93.07	83.20	81.53
GenFEND	83.65	80.88	**80.20**	**78.86**	88.23	86.01	95.48	94.66	83.03	69.79
DELL	81.25	79.54	76.65	73.52	90.79	89.87	25.67	91.26	82.02	82.21

Table 6: Results on propagation-based misinformation detection. The best results for each dataset are highlighted in **bold**.

els. On all three text datasets, the embeddings generated by LLMs improved the accuracy of MLP and EANN compared to those generated by BERT. In propagation-based misinformation detection, the embeddings generated by LLMs also led to performance improvements for GCN and BiGCN on datasets other than CED. This is because the larger scale and advanced pre-training strategies of LLMs enable their generated embeddings to more effectively capture and represent key features compared to embeddings generated by BERT and DeBERTa. Specifically, *Qwen*, which was trained on a large amount of Chinese language data, produced better embeddings than *NV* on the LTCR and Weibo COVID datasets. Conversely, on the English datasets Twitter and PHEME5, the situation was reversed.

2) **Despite the improvements observed in other datasets, the CED dataset and Twitter covid dataset presented an exception.** When GCN and BiGCN use embeddings generated by LLMs on the CED dataset, performance declined significantly. Moreover, on the Twitter COVID dataset, GCN performed best with embeddings generated by NV, while BiGCN performed worst with embeddings generated by Qwen. This could be because the specific instructions use for generating embeddings with LLMs were not well-suited for embedding comments within the propagation graph on the datasets except for the reasons about language. Additionally, previous research (Purchase, Zhao, and Mullins 2022) has indicated that GNN models can show substantial performance variations based on the types of embeddings used, which might account for the decreased effectiveness in this case.

3) **The additional text information generated by LLMs is more effective when combined with the original data.** ARG and GenFEND outperform SLMs using BERT, while DELL does not. This is because both GenFEND and ARG integrate the generated data effectively with the original data through specially designed aggregation and interaction modules, whereas DELL only sets up different expert agents to make predictions based on the generated propagation graphs. We also noticed that the distilled version of ARG (ARG-D) achieved good results. This indicates that knowledge distillation can be use to leverage the benefits of LLM-generated information without the need to deploy the LLMs.

Dicussion

This section primarily evaluates LLM-enhanced detectors to explore the potential of LLMs as enhancers in existing misinformation detection methods. The experimental results show that LLMs when use as text embedding models, provide better embeddings compared to the mainstream embedding model BERT. However, the adaptability of Graph Neural Network (GNN) models to different types of embeddings may lead to scenarios where BERT outperforms LLMs. Additionally, using LLMs for text enhancement in datasets can provide richer information to the model, and combining LLM-generated analyses of original news data with the raw data produces unexpected results. These findings highlight the potential of LLMs in enhancing existing detection models.

Conclusion

We conduct an initial evaluation of LLMs in two approaches for misinformation detection: **LLM-based detectors** and **LLM-enhanced detectors**. The results show that LLMs can achieve comparable performance in text-based misinformation detection with well-designed prompts but exhibit notably constrained capabilities in comprehending propagation structures as a detector. On the other hand, LLM-enhanced detectors outperform SLMs in most cases when generating richer textual features, producing analysis, and simulating user engagements. Our findings demonstrate that LLMs hold great potential in detecting misinformation and supporting the existing misinformation detection models. In addition to the existing use of LLMs for synthesizing data to enhance text-level enhancement, our research also highlights the potential and prospects of utilizing LLMs as detectors and for feature enhancement in misinformation detection.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (No. 62102412), China Postdoctoral Science Foundation (No. 2024M753481), and the Postdoctoral Fellowship Program of China Postdoctoral Science Foundation (No. GZC20232969).

References

au2, M. B. I.; and Katz, D. M. 2022. GPT Takes the Bar Exam. arXiv:2212.14402.

Bai, Y.; Jones, A.; Ndousse, K.; Askell, A.; Chen, A.; DasSarma, N.; Drain, D.; Fort, S.; Ganguli, D.; Henighan, T.; et al. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Bian, T.; Xiao, X.; Xu, T.; et al. 2020. Rumor detection on social media with bi-directional graph convolutional networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, 549–556.

Castillo, C.; Mendoza, M.; and Poblete, B. 2011. Information credibility on Twitter. In *Proceedings of the 20th International Conference on World Wide Web, WWW 2011, Hyderabad, India, March 28 - April 1, 2011*.

Chen, C.; and Shu, K. 2024. Combating misinformation in the age of llms: Opportunities and challenges. *AI Magazine*, 45(3): 354–368.

Chen, M.; Wei, L.; Cao, H.; Zhou, W.; Yan, Z.; and Hu, S. 2024. Propagation Structure-Semantic Transfer Learning for Robust Fake News Detection. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 227–244. Springer.

Cheng, M.; Nazarian, S.; and Bogdan, P. 2020. Vroc: Variational autoencoder-aided multi-task rumor classifier based on text. In *Proceedings of the web conference 2020*, 2892–2898.

Choi, M.; Pei, J.; Kumar, S.; Shu, C.; and Jurgens, D. 2023. Do LLMs Understand Social Knowledge? Evaluating the Sociability of Large Language Models with SocKET Benchmark. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 11370–11403.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *North American Chapter of the Association for Computational Linguistics*.

Dong, M.; Yao, L.; Wang, X.; Benatallah, B.; Sheng, Q. Z.; and Huang, H. 2018. Dual: A deep unified attention model with latent relation representations for fake news detection. In *Web Information Systems Engineering–WISE 2018: 19th International Conference, Dubai, United Arab Emirates, November 12-15, 2018, Proceedings, Part I 19*, 199–209. Springer.

Dou, Y.; Shu, K.; Xia, C.; Yu, P. S.; and Sun, L. 2021. User Preference-aware Fake News Detection. In Diaz, F.; Shah, C.; Suel, T.; Castells, P.; Jones, R.; and Sakai, T., eds., *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021*, 2051–2055. ACM.

Du, Z.; Qian, Y.; Liu, X.; Ding, M.; Qiu, J.; Yang, Z.; and Tang, J. 2021. GLM: General Language Model Pretraining with Autoregressive Blank Infilling. In *Annual Meeting of the Association for Computational Linguistics*.

Fan, L.; Li, L.; Ma, Z.; Lee, S.; Yu, H.; and Hemphill, L. 2023. A bibliometric review of large language models research from 2017 to 2023. *arXiv preprint arXiv:2304.02020*.

Faris, R.; Roberts, H.; Etling, B.; Bourassa, N.; Zuckerman, E.; and Benkler, Y. 2017. Partisanship, propaganda, and disinformation: Online media and the 2016 US presidential election. *Berkman Klein Center Research Publication*, 6.

Fisher, M.; Cox, J. W.; and Hermann, P. 2016. Pizzagate: From rumor, to hashtag, to gunfire in DC. *Washington Post*, 6: 8410–8415.

Guo, J.; Du, L.; Liu, H.; Zhou, M.; He, X.; and Han, S. 2023. Gpt4graph: Can large language models understand graph structured data? an empirical evaluation and benchmarking. *arXiv preprint arXiv:2305.15066*.

Hamed, S. K.; Ab Aziz, M. J.; and Yaakub, M. R. 2023. Fake news detection model on social media by leveraging sentiment analysis of news content and emotion analysis of users' comments. *Sensors*, 23(4): 1748.

He, P.; Liu, X.; Gao, J.; and Chen, W. 2020. DeBERTa: Decoding-enhanced BERT with Disentangled Attention. *ArXiv*, abs/2006.03654.

He, Z.; Li, C.; Zhou, F.; and Yang, Y. 2021. Rumor Detection on Social Media with Event Augmentations. In Diaz, F.; Shah, C.; Suel, T.; Castells, P.; Jones, R.; and Sakai, T., eds., *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021*, 2020–2024. ACM.

Holan, A. D. 2016. 2016 Lie of the Year: Fake news. *Politifact*.

Hu, B.; Sheng, Q.; Cao, J.; Shi, Y.; Li, Y.; Wang, D.; and 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*, volume 38, 22105–22113.

Hu, D.; Hou, X.; Du, X.; Zhou, M.; Jiang, L.; Mo, Y.; and Shi, X. 2022. VarMAE: Pre-training of Variational Masked Autoencoder for Domain-adaptive Language Understanding. In *EMNLP (Findings)*, 6276–6286.

Hu, D.; Wei, L.; Zhou, W.; Huai, X.; Han, J.; and Hu, S. 2021a. A rumor detection approach based on multi-relational propagation tree. *Journal of Computer Research and Development*, 58(7): 1395–1411.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2021b. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Huang, Y.; Shu, K.; Yu, P. S.; and Sun, L. 2024. From Creation to Clarification: ChatGPT's Journey Through the Fake News Quagmire. In *Companion Proceedings of the ACM on Web Conference 2024*, 513–516.

Huang, Y.; and Sun, L. 2023. Harnessing the Power of ChatGPT in Fake News: An In-Depth Exploration in Generation, Detection and Explanation. arXiv:2310.05046.

Ji, Z.; Yu, T.; Xu, Y.; Lee, N.; Ishii, E.; and Fung, P. 2023. Towards mitigating LLM hallucination via self reflection. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 1827–1843.

Jiang, A. Q.; Sablayrolles, A.; Mensch, A.; Bamford, C.; Chaplot, D. S.; Casas, D. d. l.; Bressand, F.; Lengyel, G.; Lample, G.; Saulnier, L.; et al. 2023. Mistral 7B. *arXiv preprint arXiv:2310.06825*.

Jwa, H.; Oh, D.; Park, K.; Kang, J. M.; and Lim, H. 2019. exbake: Automatic fake news detection model based on bidirectional encoder representations from transformers (bert). *Applied Sciences*, 9(19): 4062.

Kaliyar, R. K.; Goswami, A.; and Narang, P. 2021. FakeBERT: Fake news detection in social media with a BERT-based deep learning approach. *Multimedia tools and applications*, 80(8): 11765–11788.

Leite, J. A.; Razuvayevskaya, O.; Bontcheva, K.; and Scarton, C. 2023. Detecting Misinformation with LLM-Predicted Credibility Signals and Weak Supervision. arXiv:2309.07601.

Liu, Y.; and Wu, Y. F. B. 2018. Early Detection of Fake News on Social Media Through Propagation Path Classification with Recurrent and Convolutional Networks. In *National Conference on Artificial Intelligence*.

Lu, Y.-J.; and Li, C.-T. 2020. GCAN: Graph-aware Co-Attention Networks for Explainable Fake News Detection on Social Media. In Jurafsky, D.; Chai, J.; Schluter, N.; and Tetreault, J., eds., *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 505–514. Online: Association for Computational Linguistics.

Luvembe, A. M.; Li, W.; Li, S.; Liu, F.; and Xu, G. 2023. Dual emotion based fake news detection: A deep attention-weight update approach. *Information Processing & Management*, 60(4): 103354.

Ma, G.; Hu, C.; Ge, L.; Chen, J.; Zhang, H.; and Zhang, R. 2022. Towards Robust False Information Detection on Social Networks with Contrastive Learning. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, CIKM '22, 1441–1450. New York, NY, USA: Association for Computing Machinery. ISBN 9781450392365.

Ma, J.; Gao, W.; Mitra, P.; Kwon, S.; Jansen, B. J.; Wong, K.-F.; and Cha, M. 2016. Detecting rumors from microblogs with recurrent neural networks.

Ma, J.; Gao, W.; Wei, Z.; Lu, Y.; and Wong, K.-F. 2015. Detect rumors using time series of social context information on microblogging websites. In *Proceedings of the 24th ACM international on conference on information and knowledge management*, 1751–1754.

Ma, J.; Gao, W.; and Wong, K. 2018. Rumor Detection on Twitter with Tree-structured Recursive Neural Networks. In Gurevych, I.; and Miyao, Y., eds., *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, 1980–1989. Association for Computational Linguistics.

Ma, J.; Gao, W.; and Wong, K.-F. 2017. Detect Rumors in Microblog Posts Using Propagation Structure via Kernel Learning. In *Annual Meeting of the Association for Computational Linguistics*.

Ma, J.; Gao, W.; and Wong, K. F. 2019. Detect Rumors on Twitter by Promoting Information Campaigns with Generative Adversarial Learning.

Ma, Z.; Liu, M.; Fang, G.; and Shen, Y. 2023. LTCR: Long-Text Chinese Rumor Detection Dataset. *ArXiv*, abs/2306.07201.

Nan, Q.; Sheng, Q.; Cao, J.; Hu, B.; Wang, D.; and Li, J. 2024. Let Silence Speak: Enhancing Fake News Detection with Generated Comments from Large Language Models. *arXiv preprint arXiv:2405.16631*.

Naveed, H.; Khan, A. U.; Qiu, S.; Saqib, M.; Anwar, S.; Usman, M.; Barnes, N.; and Mian, A. 2023. A comprehensive overview of large language models. *arXiv preprint arXiv:2307.06435*.

Nguyen, V.-H.; Sugiyama, K.; Nakov, P.; and Kan, M.-Y. 2020. Fang: Leveraging social context for fake news detection using graph representation. In *Proceedings of the 29th ACM international conference on information & knowledge management*, 1165–1174.

Nori, H.; King, N.; McKinney, S. M.; Carignan, D.; and Horvitz, E. 2023. Capabilities of GPT-4 on Medical Challenge Problems. arXiv:2303.13375.

OpenAI. 2022. ChatGPT: Optimizing language models for dialogue. https://openai.com/blog/chatgpt/. Accessed: 2023-12-13.

Purchase, S.; Zhao, A.; and Mullins, R. D. 2022. Revisiting embeddings for graph neural networks. *arXiv preprint arXiv:2209.09338*.

Shu, K.; Mahudeswaran, D.; Wang, S.; Lee, D.; and Liu, H. 2018. FakeNewsNet: A Data Repository with News Content, Social Context, and Spatiotemporal Information for Studying Fake News on Social Media. *Big data*, 8 3: 171–188.

Shu, K.; Sliva, A.; Wang, S.; Tang, J.; and Liu, H. 2017. Fake news detection on social media: A data mining perspective. *ACM SIGKDD explorations newsletter*, 19(1): 22–36.

Siska, C.; Marazopoulou, K.; Ailem, M.; and Bono, J. 2024. Examining the robustness of LLM evaluation to the distributional assumptions of benchmarks. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 10406–10421. Bangkok, Thailand: Association for Computational Linguistics.

Song, C.; Shu, K.; and Wu, B. 2021. Temporally evolving graph neural network for fake news detection. *Information Processing & Management*, 58(6): 102712.

Song, C.; Tu, C.; Yang, C.; Liu, Z.; and Sun, M. 2018. CED: Credible Early Detection of Social Media Rumors. *arXiv preprint arXiv:1811.04175*.

Su, X.; Yang, J.; Wu, J.; and Zhang, Y. 2023. Mining user-aware multi-relations for fake news detection in large scale online social networks. In *Proceedings of the sixteenth ACM international conference on web search and data mining*, 51–59.

Sun, T.; Qian, Z.; Dong, S.; Li, P.; and Zhu, Q. 2022. Rumor detection on social media with graph adversarial contrastive learning. In *Proceedings of the ACM Web Conference 2022*, 2789–2797.

Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Vaibhav, V.; Annasamy, R. M.; and Hovy, E. 2019. Do sentence interactions matter? leveraging sentence level representations for fake news classification. *arXiv preprint arXiv:1910.12203*.

Vosoughi, S.; Roy, D.; and Aral, S. 2018. The spread of true and false news online. *science*, 359(6380): 1146–1151.

Wan, H.; Feng, S.; Tan, Z.; Wang, H.; Tsvetkov, Y.; and Luo, M. 2024. Dell: Generating reactions and explanations for llm-based misinformation detection. *arXiv preprint arXiv:2402.10426*.

Wang, Y.; Ma, F.; Jin, Z.; Yuan, Y.; Xun, G.; Jha, K.; Su, L.; and Gao, J. 2018. Eann: Event adversarial neural networks for multi-modal fake news detection. In *Proceedings of the 24th acm sigkdd international conference on knowledge discovery & data mining*, 849–857.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Ichter, B.; Xia, F.; Chi, E.; Le, Q.; and Zhou, D. 2023. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. arXiv:2201.11903.

Wei, L.; Hu, D.; Lai, Y.; Zhou, W.; and Hu, S. 2022a. A Unified Propagation Forest-based Framework for Fake News Detection. In *Proceedings of the 29th International Conference on Computational Linguistics, COLING 2022, Gyeongju, Republic of Korea, October 12-17, 2022*, 2769–2779. International Committee on Computational Linguistics.

Wei, L.; Hu, D.; Zhou, W.; and Hu, S. 2022b. Uncertainty-aware Propagation Structure Reconstruction for Fake News Detection. In *Proceedings of the 29th International Conference on Computational Linguistics, COLING 2022, Gyeongju, Republic of Korea, October 12-17, 2022*, 2759–2768. International Committee on Computational Linguistics.

Wei, L.; Hu, D.; Zhou, W.; Wang, X.; and Hu, S. 2022c. Modeling the uncertainty of information propagation for rumor detection: A neuro-fuzzy approach. *IEEE Transactions on Neural Networks and Learning Systems*.

Wei, L.; Hu, D.; Zhou, W.; Yue, Z.; and Hu, S. 2021. Towards Propagation Uncertainty: Edge-enhanced Bayesian Graph Convolutional Networks for Rumor Detection. In Zong, C.; Xia, F.; Li, W.; and Navigli, R., eds., *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, 3845–3854. Association for Computational Linguistics.

Wu, J.; Guo, J.; and Hooi, B. 2024. 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*, 3367–3378.

Wu, J.; and Hooi, B. 2023. DECOR: Degree-Corrected Social Graph Refinement for Fake News Detection. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2582–2593.

Yang, A.; Yang, B.; Hui, B.; Zheng, B.; Yu, B.; Zhou, C.; Li, C.; Li, C.; Liu, D.; Huang, F.; et al. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Yin, F.; Vig, J.; Laban, P.; Joty, S.; Xiong, C.; and Wu, C.-S. 2023. Did You Read the Instructions? Rethinking the Effectiveness of Task Definitions in Instruction Learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 3063–3079.

Yu, F.; Liu, Q.; Wu, S.; Wang, L.; Tan, T.; et al. 2017. A Convolutional Approach for Misinformation Identification. In *IJCAI*, 3901–3907.

Yuan, C.; Ma, Q.; Zhou, W.; Han, J.; and Hu, S. 2020. Early Detection of Fake News by Utilizing the Credibility of News, Publishers, and Users based on Weakly Supervised Learning. In *Proceedings of the 28th International Conference on Computational Linguistics*, 5444–5454.

Zhao, W. X.; Zhou, K.; Li, J.; Tang, T.; Wang, X.; Hou, Y.; Min, Y.; Zhang, B.; Zhang, J.; Dong, Z.; Du, Y.; Yang, C.; Chen, Y.; Chen, Z.; Jiang, J.; Ren, R.; Li, Y.; Tang, X.; Liu, Z.; Liu, P.; Nie, J.-Y.; and Wen, J.-R. 2023. A Survey of Large Language Models. arXiv:2303.18223.

Zheng, L.; Chiang, W.-L.; Sheng, Y.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Lin, Z.; Li, Z.; Li, D.; Xing, E. P.; Zhang, H.; Gonzalez, J. E.; and Stoica, I. 2023. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. arXiv:2306.05685.

Zubiaga, A.; Liakata, M.; and Procter, R. 2016. Learning Reporting Dynamics during Breaking News for Rumour Detection in Social Media. arXiv:1610.07363.

Appendix

Implementation Details

This section mainly explains the configurations of the models use in the evaluation.

For text-based methods, the learning rate for MLP (2-layers) and EANN is set to 2e-5, with a batch size of 32, and each training session runs for 20 epochs. We implement ARG [14] and ARG-D under the same parameter setting reported in the original papers.

For propagation-based methods, the learning rate for GCN (2-layers) and BiGCN is set to 1e-3, with a batch size of 128, and each training session runs for 35 epochs. We implement GenFEND [15] and DELL [16] with *Mistral-7B-Instruct-v0.2* on English datasets and *Qwen1.5-7B-Instruct* on Chinese datasets, and we use the same parameter setting reported in the original papers. And for the proxy tasks in DELL, we removed the **Knowledge Retrieval** task.

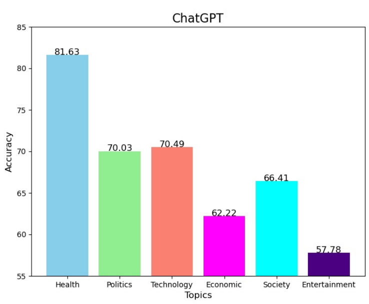
The embedding models use by SLMs (MLP, EANN, GCN, and BiGCN) and LLM-enhanced methods (ARG, GenFEND and DELL) are shown in Table 10. For methods using BERT and DeBERTa, the Chinese versions (bert-base-chinese, deberta-base-chinese [17]) are use for Chinese datasets, while the English versions (bert-base-uncased, deberta-base [18]) are use for English datasets.

All experiments were conducted on a single Tesla V100 (32GB), using PyTorch version 1.12.1 and Geometric version 2.3.1.

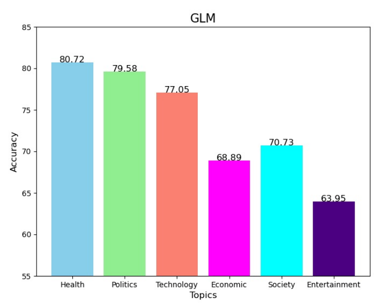
[14] https://github.com/ICTMCG/ARG
[15] https://github.com/ICTMCG/GenFEND
[16] https://github.com/whr000001/DELL
[17] https://huggingface.co/KoichiYasuoka/deberta-base-chinese
[18] https://huggingface.co/microsoft/deberta-base

Table 7: The statistics of datasets. We tallied the number of samples in eight datasets, as well as the counts of positive and negative samples (real news and fake news). Additionally, we recorded the lengths of the news and comments within the samples, along with the number of tokens consumed when inputting them into LLMs[13].

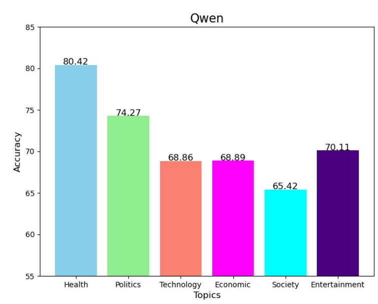
Datasets	FND23	FakeNewsNet	LTCR	Twitter	Twitter covid	CED	Weibo covid	PHEME
Number of News	1,720	23,196	2,287	1,154	400	3,387	411	5,801
Number of True News	347	17,441	1,678	575	252	1,538	264	3,829
Number of False News	1373	5,755	609	579	148	1,849	147	1972
Number of Comments	-	-	-	59,255	408,183	730,015	36,779	85,408
Max number of Comments	-	-	-	674	25,144	826	3,097	674
Average number of Comments	-	-	-	52	1,020	216	90	15
Max length of Comments(News)	394	340	4984	146	143	279	697	187
Average length of Comments(News)	106	69	207	64	58	12	36	93
Average token numbers per data entry	140	91	295	4438	78885	3456	4322	1860



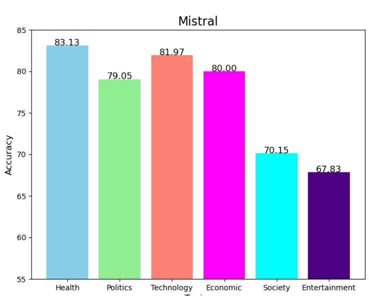
(a) ChatGPT



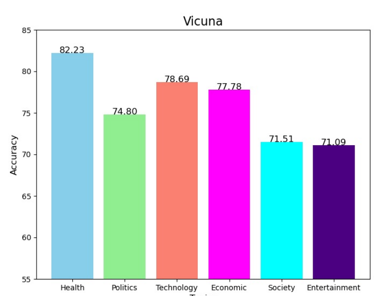
(b) GLM



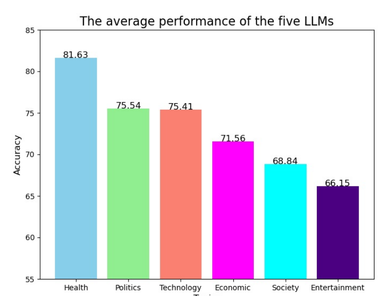
(c) Qwen



(d) Mistral



(e) Vicuna



(f) Average performance over LLMs

Figure 5: Results (%) of LLMs against different categories of news.

❶ *Vanilla prompts*

1: Given the following news, we need you to detect fake news. Please determine whether each of the news is true or false. If it is true, please only answer 1, and if it is false, please only answer 0. Do not reply with any words other than 0 or 1. For each news, you must give your answer.
Given the news `[<news>]` *, you are supposed to answer:*
2: Now you are an annotator to determine whether a given news is fake news. The news will be given in the form that 'Is [News] true news? Answer: [Answer]'. You need to give an answer in the [Answer] slot. There are two available answers that you can choose to fill the slot: 0 and 1, if the news is true, you are supposed to answer 1, otherwise, you should answer 0. You have to choose one of the two answers from 1 and 0. No need to answer any other words except 1 and 0.
Is `[<claim>]` *a true statement? Answer:* `[Answer]`
3: Now we need you to detect fake news. Below, I will provide you with a piece of news. If the news is true, please reply with 1. If the news is fake, please reply with 0. Do not reply any other words except 1 and 0.
News: `[<news>]`*, Answer:*

---

❷ *Task Prompt*

Now we need you to detect fake news. ***Fake news is defined as news that is deliberately written to mislead readers into believing false information***. Below, I will provide you with a piece of news. If the news is true, please reply with 1.If the news is fake, please reply with 0. Do not reply any other words except 1 and 0.

---

❸ *Chain-of-Thought (CoT) Prompt*

Now we need you to detect fake news. Below, I will provide you with a piece of news. If the news is true, please reply with 1.If the news is fake, please reply with 0. ***Please consider the authenticity of the news step by step before replying***. Do not reply any other words except 1 and 0.

---

❹ *Reason-Aware (RA) Prompt*

Now we need you to detect fake news. ***Here are the options for categorizing the underlying causes of fake news: [A. Emotional bias or misleading intent, B. Lack of evidence or credible sources, C. Conflicting fact, D. Informal statements, expressions, or vague language, E. insufficient supporting materials, F. Lack of context or taken out of context, G. Misinterpretation or misquotation, H. Oversimplification or exaggeration, I. Doctored images or videos, J. Other reasons]*** Below, I will provide you with a piece of news. If the news is true, please reply with 1.If the news is fake, please reply with 0. Do not reply any other words except 1 and 0."

---

❹ *Few-shot Learning (N=3)*

Now we need you to detect fake news. Below, I will provide you with a piece of news. If the news is true, please reply with 1. If the news is false, please reply with 0. Do not reply with any other words except 1 and 0.
Here are some examples of fake news examples:
News: [No Evidence Lisa Loring's Cause of Death Was COVID-19 Vaccine], Answer: [0]
News: [AP: Diamond's Cause of Death Was Heart Disease, Not COVID-19 Vaccine], Answer: [0]
News: [The U.S. Department of Defense awarded a contract for COVID-19 research in Ukraine months before the virus was known.], Answer: [0]
Here are some examples of true news examples:
News: [A photo shows Ukrainian soldiers posing in front of a NATO flag, an Azov flag, and the swastika-bearing flag once use by the Hitler Youth], Answer: [1]
News: [A video authentically shows a March 2023 incident involving a U.S. surveillance drone and Russian Su-27 fighter jets], Answer: [1]
News: [HIV is contagious.], Answer: [1]

Table 8: Prompt settings use to evaluate the performance of LLMs as detectors on text-based misinformation. For vanilla prompts, Template prompts are in **bold**, the query is in *italics*, and the input (the news to detect) is in `typewriter font`. For other prompts (such as Task Prompt), the changes compared to the Vanilla prompts are highlighted in **bold**.

❶ *Vanilla prompts*

News: Now you are an annotator to detect misinformation. You need to determine whether the given claim is misinformation. The claim will be given in the format `[Claim[id:claim]]`
Given the claim `[Claim [<id, claim>]]`*, please determine whether the claim is misinformation based on its comments and propagation structure. If the claim is true, you are supposed to answer 1; if it is false, you are supposed to answer 0. You have to choose one of the two answers, 0 or 1. No need to answer any other words except 0 or 1. Your answer is:*

Comments: Now you are an annotator to detect misinformation. You need to determine whether the given claim with comments is misinformation. The claim and its comments will be given in the format `[Claim[id:claim]], [Comments [comment_texts]]`
Given the claim `[Claim [<id, claim>]], [Comments [<id, comments>]]`*, please determine whether the claim is misinformation based on its comments. If the claim is true, you are supposed to answer 1; if it is false, you are supposed to answer 0. You have to choose one of the two answers, 0 or 1. No need to answer any other words except 0 or 1. Your answer is:*

Propagation Structure: Now you are an annotator to detect misinformation. You need to determine whether the given claim with comments and propagation structure composed of comments is misinformation. The claim, its comments, and the propagation structure composed of comments will be given in the format `[Claim[id:claim]], [Comments [id:comment_texts]], [propagation structure: ['id1 replied to id2',...]]`. `[id:comment_text]` means the ID and text of the comments on the tweet, and `'id1 replied to id2'` means the comment with ID `id1` is a reply to the comment with ID `id2`.
Given the claim `[Claim [<id, claim>]], [Comments [<id, comments>]]`, *and* `[propagation structure [<propagation structure>]]`*, please determine whether the claim is misinformation based on its comments and propagation structure. If the claim is true, you are supposed to answer 1; if it is false, you are supposed to answer 0. You have to choose one of the two answers, 0 or 1. No need to answer any other words except 0 or 1. Your answer is:*

❷ *Refine Prompt*

Now you are an annotator to detect misinformation. You need to determine whether the given claim with comments and propagation structure composed of comments is misinformation. The claim, its comments, and the propagation structure composed of comments will be given in the format [Claim[id: claim]], [Comments [id:comment_texts]],[propagation structure: ['id1 replied to id2',...]]. [id:comment_text] means the ID and text of the comments on the tweet, and 'id1 replied to id2' means the comment with ID id1 is a reply to the comment with ID id2.
Due to the noise in the comments and dissemination structure of the tweet, before you make a judgment, you need to remove comments that are irrelevant to the authenticity of the tweet and optimize the propagation structure based on the remaining replies, and then judge whether the tweet is a rumor based on the propagation structure optimized by yourself.

❸ *Format Graph Input*

Now you are an annotator to detect rumors. You need to determine whether the given claim is misinformation based on its propagation graph. The claim will be given in the form `[ID: claim]` and its propagation graph will be given in the form:
`Nodes: ['id1:content1', 'id2:content2', ...]`
`Edges: [[All parent nodes' IDs], [All child nodes' IDs]]`
`Root node: ['id of the root node: the content of the root node']`

Table 9: Prompt settings use to evaluate the performance of LLMs as detectors on propagation-based misinformation. We only present one Vanilla prompt and its variants.

Methods	**Text embedding methods**	**Feature dimension**
SLMs(SLMs w/BERT) & Text-level enhancement methods	bert-base-uncased bert-base-chinese	768 768
SLMs w/DeBERTa	deberta-base deberta-base-chinese	768 768
SLMs w/NV	NV-Embed-v2[19]	4096
SLMs w/Qwen	gte-Qwen2-1.5B-instruct[20]	1536

Table 10: Text embedding models use by different methods.

❶ *Failure to effectively integrate analysis results*

Q: Now we need you to detect fake news. Below, I will provide you with a piece of news. If the news is true, please reply with 1. If the news is fake, please reply with 0. ***Please consider the authenticity of the news and provide your reasons along with your answer.*.**
News: [<Joe Bidden closed schools during the COVID-19 pandemic>], Answer:
A: Joe Bidden did not close schools during the COVID-19 pandemic. Schools were closed as a result of government-issued shutdown orders to migrate the spread of the virus. Answer: **True**.
Label: Fake

---

❷ *Wrong analytical process*

Q: Now we need you to detect fake news. Below, I will provide you with a piece of news. If the news is true, please reply with 1. If the news is fake, please reply with 0. ***Please consider the authenticity of the news and provide your reasons along with your answer.*.**
News: [<COVID Survivor Claire Bridges Sets Record Straight on Misleading Tweet About Her>], Answer:
A: Since the news is focuse on correcting information and suggests that there was a tweet circulating that contained inaccurate information about Claire Bridges, she has since clarified the facts to correct the record, which makes the statement true. Answer: **True**.
Label: Fake

---

❸ *Unrelated analyses*

Q: Now we need you to detect fake news. Below, I will provide you with a piece of news. If the news is true, please reply with 1. If the news is fake, please reply with 0. ***Please consider the authenticity of the news and provide your reasons along with your answer.*.**
News: [<A music video features Ukrainian President Volodymyr Zelenskyy dancing in high heels, a cropped top, and skin-tight pants with three other men.>], Answer:
A: This statement is about Zelenskyy dancing in high heels, not a newsworthy event. This may be a clickbait headline and seems to be a fabrication. Answer: **Fake**.
Label: True

Table 11: Three main factors contributing to worse performance of LLMs: (a) failure to effectively integrate analysis results, (b) wrong analytical process, and (c) unrelated analyses