Exploring the Truth with Dialogue: Dual Large Language Model Interaction Cooperation and Multi-view Semantic Fusion Network for Fake News Detection

Anonymous ACL submission

Abstract

The widespread dissemination of fake news poses a significant threat to social trust and individual decision-making, necessitating advanced fake news detection technologies. Although integrating small language models (SLMs) with large language models (LLMs) has shown promise in detecting fake news, existing fake news detection methods with LLMs exploit large language models to generate extra knowledge of the social context for fake news detection. However, the LLMs themselves suffer from the hallucinations - generating plausible yet factually incorrect content. In addition, the SLMs of current methods focus on data consistency rather than data diversity when integrating multivariate information, resulting in incomplete information fusion. To address these challenges, we propose a novel fake news detection framework DLLM-MVSFN that combines a dual large language model interaction cooperation module and a multi-view semantic fusion network. DLLM-MVSFN leverages an interactive dialogue between two LLMs to generate comprehensive summaries of news events. Then a multi-view semantic fusion network is proposed to effectively integrate information from news content, LLMs summaries, and user comments for fake news detection. The experimental results show that our proposed DLLM-MVSFN outperforms existing baselines in multiple public datasets, achieving higher accuracy and F1 scores.

1 Introduction

Fake news proliferation is one of the most significant challenges facing modern society. The rapid advancement of the internet and social media platforms has dramatically accelerated the circulation of information, enabling fake news to spread at unprecedented rates (Tasnim et al., 2020). This phenomenon seriously threatens public opinion, social stability, and democratic systems. For instance, during the 2016 U.S. presidential election,

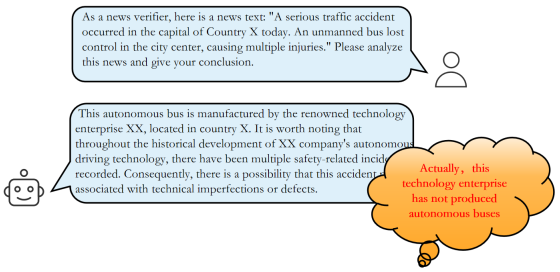

Figure 1: An example of LLMs hallucinations for fake news detection.

the rampant dissemination of fake news not only undermined public trust in science and genuine journalism but also altered societal consensus on several critical issues (Olan et al., 2022).

Traditional news verification methods, such as fact-checking (Yang et al., 2024b) and examination of dissemination patterns (Vosoughi et al., 2018), have struggled to keep pace with the exponential growth of information. Consequently, automatic fake news detection (FND) has emerged as a key research focus to mitigate the adverse impacts caused by false information. Pre-trained small language models (SLMs) like BERT and RoBERTa have proven effective for fake news detection(Angizeh and Keyvanpour, 2024; Devlin et al., 2019; Nan et al., 2021a,b; Zhu et al., 2022b; Wang et al., 2018). Fine-tuning SLMs can integrate context information more effectively (Hu et al., 2023) for fake news detection. However, SLMs-based methods lack social context knowledge, limiting their performance improvements. Moreover, the SLMs of current methods focus on data consistency rather than data diversity when integrating multivariate information, leading to incomplete information fusion and inadequate adaptability in complex environments (Wu et al., 2021; Yang et al., 2024a).

The emergence of large language models (LLMs) (OpenAI, 2022; Kalyan, 2024) provides an option to supplement social context knowledge for

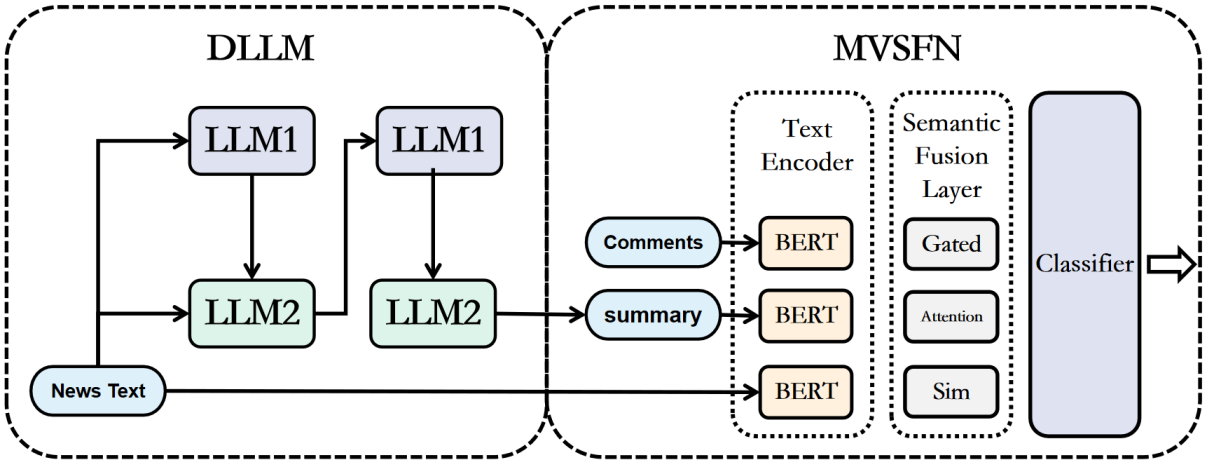

Figure 2: The architecture of DLLM-MVSFN. DLLM-MVSFN consists of a dual large language model interaction cooperation module and a multi-view semantic fusion network. The dual large language model interaction cooperation module engages in a simulated expert dialogue with two language models to relieve hallucinations and deeply explore the social context knowledge of news. The multi-view semantic fusion network employs three fusion techniques to effectively integrate information from news content, LLMs summaries, and user comments for fake news detection.

SLMs-based fake news detection methods. LLMs leverage extensive training on vast corpora, equipping them with rich knowledge bases and robust generalization capabilities (Grosse et al., 2023). This enables them to understand and analyze information within broader contexts. Therefore, existing fake news detection methods (Chen et al., 2024; Ma et al., 2024) exploit large language models to generate extra knowledge of the social context for fake news detection. However, the LLMs suffer from hallucinations problem(Huang et al., 2023)—content that appears plausible but is factually incorrect or misleading, which will introduce noise information for fake news detection. As shown in Figure 1, LLMs generate extra knowledge of the social context for fake news detection, but the generated knowledge appears plausible and is incorrect factually.

To address these challenges, we propose a novel fake news detection framework, DLLM-MVSFN, consisting of a dual large language model interaction cooperation module and a multi-view semantic fusion network. DLLM-MVSFN leverages an interactive dialogue between two LLMs to generate comprehensive summaries of news events. Then a multi-view semantic fusion network is proposed to effectively integrate information from news content, LLMs summaries, and user comments for fake news detection. Specifically, the dual large language model interaction cooperation module engages in a simulated expert dialogue with two language models, one model acts as the questioner or critic, and the other serves as the responder or analyzer, to relieve hallucinations and deeply explore the social context knowledge of news. The multi-view semantic fusion network employs three fusion techniques, i.e., similarity-weighted fusion, attention-weighted fusion, and gated attention fusion, to effectively integrate information from news content, LLMs summaries, and user comments for fake news detection. The experimental results show that our proposed DLLM-MVSFN outperforms existing baselines in multiple public datasets, achieving higher accuracy and F1 scores. The main contributions are summarized:

- To the best of our knowledge, we are the first to propose a novel dual LLM interaction mechanism in which one LMM acts as a questioner/challenger and the other plays the role of answerer/analyzer to relieve LLMs' hallucinations and generate in-depth and accurate social context knowledge for fake news detection.
- We explore a multi-view semantic fusion network that employs similarity-weighted, attention-weighted, and gated attention fusion to effectively integrate information from news content, LLMs summaries, and user comments for fake news detection.

- The experimental results on multiple public datasets show that our proposed fake news detection framework DLMM-MVSFN achieves higher accuracy and F1 scores than existing baselines.

2 DLLM-MVSFN

As shown in Figure 2, DLLM-MVFSN consists of a dual large language model interaction cooperation module and a multi-view semantic fusion network. The dual large language model interaction cooperation module engages in a simulated expert dialogue with two language models, one model acts as the questioner or critic, and the other serves as the responder or analyzer, to relieve hallucinations and deeply explore the social context knowledge of news. The multi-view semantic fusion network employs three fusion techniques, i.e., similarity-weighted fusion, attention-weighted fusion, and gated attention fusion, to effectively integrate information from news content, LLMs summaries, and user comments for fake news detection.

2.1 Dual LLM Interactive Cooperation Mechanism

As illustrated in Figure 3, our approach leverages two large language models (LLMs) to collaboratively perform the roles of **questioner/critic** and **responder/summarizer**. This framework is inspired by expert discussions, where diverse perspectives and analytical approaches are employed to address complex problems. The interaction unfolds in four sequential stages: **inquiry**, **response**, **critique**, and **synthesis**.

Inquiry In the initial stage, **LLM1** processes the news text and generates probing questions aimed at evaluating the credibility and authenticity of the article. By employing prompt engineering (Vatsal and Dubey, 2024), LLM1 formulates diverse and critical questions that lay the foundation for subsequent analysis.

Response Building upon the questions posed by LLM1, **LLM2** analyzes the same news content along with LLM1's outputs, providing detailed and well-reasoned responses. This ensures comprehensive addressing of the questions with an emphasis on factual accuracy and logical coherence.

Critique In this stage, **LLM1** assumes the role of critic, rigorously evaluating the quality, depth, and relevance of the responses provided by LLM2. The critique process identifies gaps, inconsistencies, or areas requiring further clarification, thus enhancing the robustness of the analysis.

Synthesis Finally, **LLM2** synthesizes all gathered information—including the original news text, user comments, questions, and critiques—into a coherent and comprehensive summary. This output captures the core essence of the news article while reflecting the diverse perspectives that emerged during the iterative interaction.

The dual LLM interactive cooperation mechanism challenges each other's assumptions, mitigates biases, and uncovers blind spots through the above-mentioned process of inquiry, response, critique, and Synthesis, thereby enhancing the interpretability and reliability of the final analysis. The mechanism draws strength from the inherent differences between the two models, arising from variations in their training corpora, architectures, and methodologies (Pimentel et al., 2024). For instance, one model might excel in factual recall due to extensive training on encyclopedic data, while the other demonstrates nuanced inferential reasoning derived from conversational datasets (Lu et al., 2024). The differences between the two LLMs provide complementary knowledge bases and analytical strategies, thus the dual LLM interactive cooperation mechanism ensures a nuanced and accurate representation of the news and the social context knowledge enriched by multiple layers of analysis and dialogue.

2.2 Multi-view Semantic Fusion Network(MVSFN)

As shown in Figure 4, our proposed Multi-view Semantic Fusion Network (MVSFN) comprises a text encoder, a semantic fusion layer, and a classifier. The text encoder encodes news-related text content, comments, and analytical summaries of social context knowledge obtained by the interaction of the two LLMs into a semantic representation. The semantic fusion layer employs three fusion techniques, i.e., similarity-weighted fusion, attention-weighted fusion, and gated attention fusion, to effectively integrate information from news content, LLMs summaries, and user comments. Finally, the integrated information is fed to a classifier to detect fake news.

2.2.1 Text Encoder

To capture the semantic information of news text, user comments, and social context knowledge

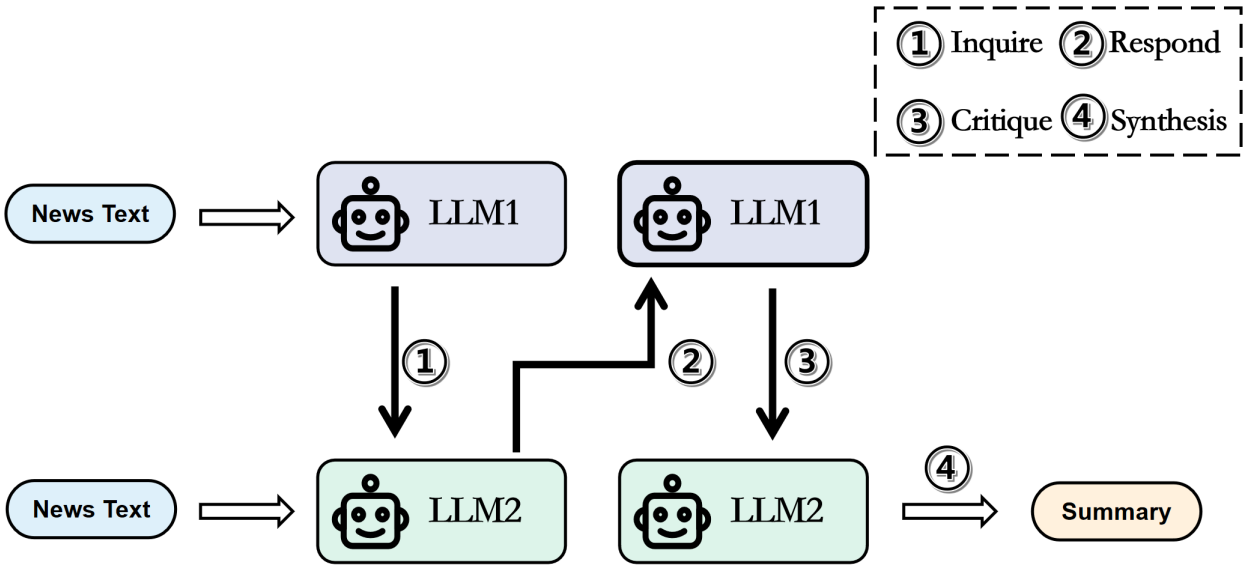

Figure 3: The illustration of Dual LLM Interactive Cooperation Mechanism. One LLM focuses on generating critical questions and evaluations, while the other delivers detailed responses and synthesizes a holistic summary.

summaries generated by LLMs, we exploit a pre-trained BERT (Devlin et al., 2019) to embed the news text, user comments, and social context knowledge summaries.The representation of news text $\mathbf{X}_{\text{news}}$, user comments $\mathbf{X}_{\text{comments}}$, and social context knowledge $\mathbf{X}_{\text{summary}}$ is formula as follows:

$$\mathbf{X}_i = \text{BERT}(\text{input}_i), \quad i \in \{\text{news}, \text{comments}, \text{summary}\} \tag{1}$$

2.2.2 Semantic Fusion Layer

To effectively integrate the semantic information of news text, user comments, and social context knowledge summaries, the semantic fusion layer employs three fusion techniques, i.e., similarity-weighted fusion, attention-weighted fusion, and gated attention fusion, to integrate information from different perspectives.

Similarity-weighted Fusion: the similarity-weighted mechanism utils cosine similarity to determine whether the information is sematic similar and captures the similar semantic information defined by cosine from different sources. it computes pairwise cosine similarities among $\mathbf{X}_{\text{news}}$, $\mathbf{X}_{\text{comments}}$, and $\mathbf{X}_{\text{summary}}$ to derive attention weights. The fused representation is computed as follows:

$$s_{ij} = \frac{\mathbf{X}_i \cdot \mathbf{X}_j}{\|\mathbf{X}_i\| \|\mathbf{X}_j\|}, \tag{2}$$

$$w_i = \text{softmax}\left(\sum_j s_{ij}\right), \tag{3}$$

$$\mathbf{X}_{\text{sim}} = \sum_i w_i \mathbf{X}_i \tag{4}$$

where s_{ij} represents the similarity score between source i and j, $i, j \in \{\text{news}, \text{comments}, \text{summary}\}$. w_i are the normalized attention weights obtained by applying the $softmax$ function to the sum of similarities for each source i.

Attention-weighted Fusion: the attention-weighted fusion assigns adaptive weights to the three source, dynamically adjusting their contributions. The formula of the attention-weighted fusion is as follows (Vaswani et al., 2017):

$$a_i = \text{Softmax}(\text{Linear}(\mathbf{X}_i)), \tag{5}$$

$$\mathbf{X}_{\text{att}} = \sum_i a_i \mathbf{X}_i \tag{6}$$

where a_i denotes the attention score for the i-th source.

Gated attention Fusion: the gated attention fusion exploits a learnable gate to weights each source and enable dynamic adjustment of its influence. The computing of the gated attention fusion

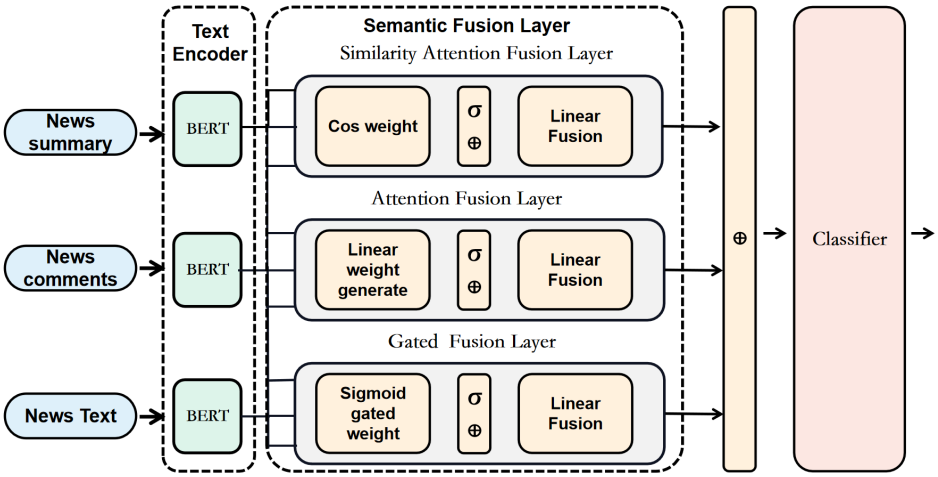

Figure 4: The architecture of MVSFN. MVSFN comprises a text encoder, a semantic fusion layer, and a classifier.

is given by (Wang et al., 2025):

$$g_i = \sigma(\text{Linear}(\mathbf{X}_i)), \tag{7}$$

$$\mathbf{X}_{\text{gated}} = \sum_i g_i \mathbf{X}_i \tag{8}$$

where g_i is the gate value for the i-th source, computed using a sigmoid activation function.

Thus, we exploit three fusion techniques to compute the semantic representation $\mathbf{X}_{\text{sim}}$, $\mathbf{X}_{\text{att}}$, and $\mathbf{X}_{\text{gated}}$ of the fused news text, user comments, and LLM-generated social context summaries. However, the direct cascading of three fused semantic representations will result in too large a dimension and thus reduce the model performance. Therefore, we use three linear layers to reduce the dimension of the fused semantic representation and then cascade it. The formalization is as follows:

$$\begin{aligned} \mathbf{Z}_i &= \text{Linear}(\mathrm{X}_i), \\ i &\in \{\text{sim}, \text{att}, \text{gated}\}, \end{aligned} \tag{9}$$

$$\mathbf{Z} = [\mathbf{Z}_{\text{sim}}, \mathbf{Z}_{\text{att}}, \mathbf{Z}_{\text{gated}}]. \tag{10}$$

2.2.3 Classifier

Finally, we fed the cascaded fusion semantic representation $\mathbf{Z}$ into a feedforward neural network with an activation function $softmax$ to detect fake news. The calculation process is as follows:

$$\hat{y} = \text{Softmax}(\text{MLP}(\mathbf{Z})), \tag{11}$$

where $\hat{y}$ represents the probabilities of being predicted as fake news. MLP denotes a multi-layer perceptron.

3 Experiments

In this section, we conduct extensive experiments to verify the performance of the proposed DLLM-MVSFN framework on fake news detection tasks. The reproducible codes and datasets used in this paper are available on GitHub.

3.1 Datasets

Our experiments were conducted on two widely recognized datasets for fake news detection: Weibo21(Nan et al., 2021a) and GossipCop(Shu et al., 2018). To mimic real-world conditions, the datasets were divided according to their temporal sequence. Each dataset contains labeled instances of real and fake news articles, providing a comprehensive benchmark for evaluating the performance of our proposed Multi-view Semantic Fusion Network(MVSFN). Table 1 summarizes the dataset statistics.

3.2 Comparison methods

We compared our proposed DLLM-MVSFN model with the following baselines: (1)**BERT**(Devlin et al., 2019): A pre-trained language model fine-tuned for fake news detection. (2)**ENDEF**(Zhu et al., 2022a): Removes entity bias and extracts generalizable features. (3)**EANN-text**(Wang et al., 2018): Focuses on event-invariant representations using text. (4)**ARG**(Hu et al., 2023): Utilizes Adaptive Reasoning Guidance for complementary insights. (5)**dEFEND with GenFEND**(Shu et al., 2019; Nan et al., 2024): Incorporates sentence-comment co-attention. (6)**DualEmo with GenFEND**(Guo et al., 2019): Considers publisher and

Dataset	Training Set		Validation Set		Test Set	
	Real News	Fake News	Real News	Fake News	Real News	Fake News
Weibo21	2989	3148	491	872	1065	252
GossipCop	3118	3164	1102	263	1079	1052

Table 1: Statistics of Weibo21 and GossipCop

Model	Weibo21				Gossipcop			
	Acc	F1	F1 - r	F1 - f	Acc	F1	F1 - r	F1 - f
BERT	0.782	0.781	0.805	0.776	0.826	0.807	0.867	0.748
ENDEF	0.772	0.770	0.787	0.753	0.846	0.830	0.883	0.795
EANN - text	0.724	0.721	0.749	0.747	0.890	0.835	0.873	0.763
ARG	0.786	0.784	0.804	0.764	0.878	0.790	0.926	0.653
dEFEND(G)	0.819	0.810	0.830	0.799	0.891	0.890	0.913	0.851
DualEmo(G)	0.808	0.808	0.812	0.810	0.914	0.900	0.936	0.869
CAS - FEND(G)	0.820	0.822	0.831	**0.811**	**0.939**	0.925	**0.948**	0.894
DLLM - MVSFN	**0.881**	**0.846**	**0.919**	0.7726	0.934	**0.934**	0.937	**0.932**

Table 2: Fake news detection results of different methods on Weibo21 and Gossipcop Datasets. (G) indicates models enhanced with the GenFEND method. Acc denotes Accuracy, F1 represents F1 - score, F1 - r indicates F1 - real, and F1 - f denotes F1 - fake.

social emotions for detection. (7)**CAS-FEND(tea) with GenFEND**(Nan et al., 2023): Leverages user comments for semantic and emotional analysis.

3.3 Experimental Setup

We evaluated the models using standard metrics such as accuracy, F1-score, F1-real, and F1-fake. In the DLLM, **LLM1** is chosen to be the GLM-4-AIR model(GLM et al., 2024), while **LLM2** is selected as the Qwen-Plus model(Bai et al., 2023). The temperature parameter for Qwen-Plus was set to 0.7, with a nucleus sampling probability (top_p) of 0.8. Conversely, GLM-4-AIR was configured with a higher temperature of 0.95 and a top_p value of 0.7. We utilized Chinese-RoBERTa-WWM-Ext for the weibo21 dataset and BERT-base-uncased for the GossipCop dataset in the MVSFN. The AdamW optimizer was selected across all models, with a uniform learning rate of 2×10^{-5}. We employed categorical cross-entropy as the loss function.

3.4 Fake news detection performance

Table 2 shows the comparison of our proposed framework with the baselines. We mark the best results in each column on the table. As shown in Table 2, on the whole, our proposed DLLM-MVSFN outperforms all the state-of-the-art approaches on both datasets. Specifically, our framework achieves an $F1$ score of 84.6% and 93.4%, respectively, increasing by 2.4% and 0.9% compared with the best baseline.

On Weibo21, we can observe that the performance of baseline CAS-FEND(G) is better than our proposed DLLM-MVSFN. The reason behind this is that there is relatively little fake news in the test data set of Weibo21, where the predicted wrong fake news label has a greater impact on the indicator. In addition, we can observe that the baseline CAS-FEND(G) outperforms the proposed DLLM-MVSFN on the Acc, F1-r, and F1-f on Gossipcop. We believe that it is because CAS-FEND(G) uses GPT to generate user comments, while our method uses two large language models pre-trained on Chinese datasets environment, whose ability of understanding english is weaker than GPT.

3.5 Ablation study

In order to study the contribution of each component in the MVSFN unit to fake news detection, we conduct ablation experiments in this part. The ablation experiments include the following six variants of the MVSFN unit:

- without Similarity-weighted fusion: Remove the Similarity-weighted fusion in the semantic fusion layer, and just use Attention-weighted Fusion and Gated attention Fusion for fusion.
- without Attention-weighted Fusion: Elimi-

Variant	Weibo21 Dataset				Gossipcop Dataset			
	Acc	F1	F1-r	F1-f	Acc	F1	F1-r	F1-f
Full DLLM-MVSFN	**0.8805**	**0.8458**	**0.919**	**0.7726**	**0.9343**	**0.9342**	**0.9367**	**0.9318**
Without Similarity-weighted fusion	0.8576	0.828	0.8994	0.7563	0.9228	0.9227	0.9218	0.9237
Without Attention-weighted fusion	0.8543	0.8163	0.9000	0.7327	0.9249	0.9249	0.9261	0.9237
Without Gated attention fusion	0.8216	0.7982	0.8669	0.7295	0.9267	0.9267	0.9287	0.9247
Only Similarity-weighted fusion	0.8642	0.8341	0.9047	0.7635	0.9127	0.9125	0.9165	0.9085
Only Attention-weighted fusion	0.7872	0.7559	0.8434	0.6684	0.9188	0.9188	0.9198	0.9178
Only Gated attention fusion	0.8151	0.7962	0.8582	0.7341	0.9076	0.9076	0.9084	0.9068

Table 3: Results of ablation study. Acc denotes Accuracy, F1 represents F1-score, F1-r indicates F1-real, and F1-f denotes F1-fake.

Variant	Accuracy	F1-score	Dataset
LLM1	0.6174	0.6546	Weibo21
LLM2	0.6513	0.6865	Weibo21
DLLM	**0.7202**	**0.708**	Weibo21
LLM1	0.5900	0.5900	Gossipcop
LLM2	0.5400	0.5455	Gossipcop
DLLM	**0.6160**	**0.6061**	Gossipcop

Table 4: LLM zero-shot Results on Dataset

nate the Attention-weighted fusion within the semantic fusion layer. Instead, rely solely on the Similarity-weighted mechanism and Gated attention fusion for the process of semantic integration.

- without Gated attention Fusion: Exclude the Gated attention fusion used for fusion in the semantic fusion layer, and conduct fusion solely via Similarity-weighted fusion and Attention-weighted fusion.
- only Similarity-weighted fusion: Only retain the Similarity-weighted fusion as the single fusion method in the semantic fusion layer.
- only Attention-weighted fusion: Only retain the Attention-weighted fusion as the single fusion method in the semantic fusion layer.
- Only Gated attention fusion: Only retain the Gated Fusion as the single fusion method in the semantic fusion layer.

Table 3 presents the ablation study results. Removing any fusion mechanism from DLLM-MVSFN results in a noticeable performance drop, highlighting the importance of each component in the network. The full DLLM-MVSFN consistently outperforms its variants, confirming the effectiveness of its design.

In addition, to verify the role of the dual LLMs interactive cooperation mechanism, we exploit any model of the dual language model and the interactive cooperation mechanism of the dual LLMs to detect fake news. The experimental results are shown in Table 4, which shows that the dual LLMs interactive cooperation mechanism can effectively improve the performance of fake news detection.

4 Related Work

In this section, we introduce fake news detection methods from three aspects: content-based, propagation-based, and knowledge-based. Compared with external knowledge, news content and its propagation structure in social media are easier to obtain. Therefore, early fake news detection research mainly focused on content and propagation structure, i.e., content-based and propagation-based methods. However, since rumor makers can easily manipulate the content and propagation structure of news, the detection method becomes ineffective. Some researchers introduce external knowledge for fake news detection, i.e., knowledge-based methods.

4.1 Content-Based Fake News Detection

Content-based fake news detection methods primarily rely on analyzing the textual content of news articles. These methods extract features from news texts and utilize machine learning or deep learning models for classification to ascertain the authenticity of the news. The core of content-based methods lies in analyzing the semantics and sentiment information within news articles to identify contradic-

tions, inconsistencies, or inflammatory language commonly found in fake news. Sentiment analysis techniques are widely applied as fake news often contains strong emotional biases or inflammatory language, which can be identified through such analyses, thereby aiding in the detection of fake news (Angizeh and Keyvanpour, 2024; Xiao et al., 2024). Studies like (Xu et al., 2025) and (Guo et al., 2019) also emphasize the importance of sentiment analysis in detecting fake news. However, these approaches have potential inaccuracy when dealing with complex language use, including satire or ambiguous expressions, which can obscure the true nature of the news, affecting overall detection accuracy. Moreover, sophisticatedly crafted fake news designed to evade detection poses additional challenges.

4.2 Propagation-Based Fake News Detection

Propagation-based fake news detection methods focus on analyzing the dissemination pathways and diffusion processes of news. Given that fake news often spreads rapidly through social networks or other dissemination platforms, particularly when it contains surprising or inflammatory content, examining these propagation patterns can aid in distinguishing between genuine and false information. By employing social network analysis techniques, researchers have identified notable differences in the topological structures of dissemination networks for fake versus real news. These distinctions offer valuable features for the early detection of fake news (Nawaz et al., 2024; Pal and Chua, 2019). Consequently, the analysis of news propagation patterns provides a critical tool for identifying fake news (Song et al., 2022; Zhang et al., 2025; Tu et al., 2021). However, Propagation-based methods depend heavily on the availability and accessibility of comprehensive dissemination data. Issues such as incomplete or biased data may lead to inaccurate detection results. Furthermore, malicious actors might manipulate dissemination patterns to mimic those of real news, thereby undermining the effectiveness of these detection methods.

4.3 Knowledge-Based Fake News Detection

Rumor spreaders can easily manipulate news content and its propagation structure. Thus, some researchers use external knowledge for fake news detection. Knowledge-based methods verify news information against external knowledge bases. Knowledge graphs are commonly used, matching news entities with real-world ones (Ma et al., 2023; Mayank et al., 2021; Fu et al., 2023; Nguyen and Do, 2024).

For instance, Mayank et al. proposed DEAP-FAKED, a knowledge graph-based method. It combines NLP and tensor decomposition, encodes news and embeds entities separately, and reduces biases in preprocessing for higher accuracy. Fu et al. introduced KG-MFEND, an efficient multi-domain model. It builds a new knowledge graph, enriches background knowledge, addresses embedding and noise issues, and uses label smoothing for strong generalization.

Recently, with large language models (LLMs) rich in knowledge, some LLM-based fake news detection methods have appeared. They enhance fact-checking by matching news with knowledge graph entities (Hu et al., 2023; Nan et al., 2024) and use reasoning to analyze and verify news semantics. However, LLMs have the hallucination problem (Ibrishimova and Li, 2020; Huang et al., 2023), introducing noise to detection.

5 Conclusion and Future Work

In this paper, we proposed a novel fake news detection framework DLLM-MVSFN to relieve the problem of LLMs' hallucinations and incomplete information fusion of SLMs in existing knowledge-based fake news detection methods. DLLM-MVSFN leverages an interactive dialogue between two LLMs to relieve hallucinations and deeply explore the social context knowledge of news. Then a multi-view semantic fusion network with similarity-weighted, attention-weighted, and gated attention fusion is explored to effectively integrate information from news content, LLMs summaries, and user comments for fake news detection. The experimental results on multiple public datasets show that our proposed fake news detection framework DLLM-MVSFN achieves higher accuracy and F1 scores than existing baselines.

In the future, we will focus on sensitive word management, prompt engineering, interaction efficiency, and data source diversification to improve the performance of knowledge-based methods.

Limitations

Despite its promising achievements, the DLLM-MVSFN framework has several limitations. The experimental results reveal that the F1 score for fake news (F1fake) is lower than that for real news

(F1real), highlighting a disparity in the model's effectiveness between identifying fake and real news. This discrepancy can be attributed to the inherent diversity and complexity of fake news, where some false information is challenging to accurately capture using current feature extraction and fusion methods.

Moreover, while the model excels at integrating multi-source information, it may still struggle to fully explore and analyze intricate semantic relationships and subtle false clues within fake news. Additionally, the reliance on large-scale pre-trained language models (LLMs) necessitates substantial computational resources, which could limit practical applications.

Limitations of Large Language Models:

Sensitive words within content pose a significant challenge to LLMs, impacting output accuracy. Fake news often includes more of these sensitive terms designed to attract attention or mislead readers, complicating accurate identification. Furthermore, prompt design constraints can lead LLMs to generate irrelevant information, such as suggesting users "refer to a specific website for more details," which detracts from core analysis tasks. The dual-model interaction process also consumes considerable tokens and time, reducing efficiency and increasing computational costs.

Challenges with Small Language Models: Integrating user comments into small language models (SLMs) enhances detection capabilities but is less effective during the initial release phase of news articles when sufficient user feedback has not yet accumulated. This reliance on user-generated content limits early-stage detection efficacy.

Acknowledgments

I would like to express my sincere gratitude for the invaluable assistance provided by several AI tools during the course of this research. Firstly, special thanks to Tongyi Qianwen for its exceptional support in automatic code completion, which greatly enhanced the efficiency of my coding process. Furthermore, I am deeply thankful for the insights provided by Qwen, Kimi, and DouBao, which were instrumental in the literature search and paper refinement. These tools have played a pivotal role in refining the scope and depth of this study, contributing significantly to its completion.

Without the aid of these advanced AI technologies, the preparation and finalization of this manuscript would have been considerably more challenging. Their contributions have been crucial in elevating the quality of this work.

References

Leila Behboudi Angizeh and Mohammad Reza Keyvanpour. 2024. Detecting fake news using advanced language models: Bert and roberta. In *2024 10th International Conference on Web Research (ICWR)*, pages 46–52.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Hao Chen, Hui Guo, Baochen Hu, Shu Hu, Jinrong Hu, Siwei Lyu, Xi Wu, and Xin Wang. 2024. A self-learning multimodal approach for fake news detection.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics*.

Lifang Fu, Huanxin Peng, and Shuai Liu. 2023. Kgmfend: an efficient knowledge graph-based model for multi-domain fake news detection. *The Journal of Supercomputing*, 79(16):18417–18444.

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, Hao Yu, Hongning Wang, Jiadai Sun, Jiajie Zhang, Jiale Cheng, Jiayi Gui, Jie Tang, Jing Zhang, Juanzi Li, Lei Zhao, Lindong Wu, Lucen Zhong, Mingdao Liu, Minlie Huang, Peng Zhang, Qinkai Zheng, Rui Lu, Shuaiqi Duan, Shudan Zhang, Shulin Cao, Shuxun Yang, Weng Lam Tam, Wenyi Zhao, Xiao Liu, Xiao Xia, Xiaohan Zhang, Xiaotao Gu, Xin Lv, Xinghan Liu, Xinyi Liu, Xinyue Yang, Xixuan Song, Xunkai Zhang, Yifan An, Yifan Xu, Yilin Niu, Yuantao Yang, Yueyan Li, Yushi Bai, Yuxiao Dong, Zehan Qi, Zhaoyu Wang, Zhen Yang, Zhengxiao Du, Zhenyu Hou, and Zihan Wang. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *Preprint*, arXiv:2406.12793.

Roger Baker Grosse, Juhan Bae, Cem Anil, Nelson Elhage, Alex Tamkin, Amirhossein Tajdini, Benoit Steiner, Dustin Li, Esin Durmus, Ethan Perez, Evan

Hubinger, Kamil.e Lukovsiut.e, Karina Nguyen, Nicholas Joseph, Sam McCandlish, Jared Kaplan, and Sam Bowman. 2023. Studying large language model generalization with influence functions. *ArXiv*, abs/2308.03296.

Chuan Guo, Juan Cao, Xueyao Zhang, Kai Shu, and Miao Yu. 2019. Exploiting emotions for fake news detection on social media.

Beizhe Hu, Qiang Sheng, Juan Cao, Yuhui Shi, Yang Li, Danding Wang, and Peng Qi. 2023. Bad actor, good advisor: Exploring the role of large language models in fake news detection. In *AAAI Conference on Artificial Intelligence*.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ArXiv*, abs/2311.05232.

Marina Danchovsky Ibrishimova and Kin Fun Li. 2020. A machine learning approach to fake news detection using knowledge verification and natural language processing. In *Advances in Intelligent Networking and Collaborative Systems*, pages 223–234, Cham. Springer International Publishing.

Katikapalli Subramanyam Kalyan. 2024. A survey of gpt-3 family large language models including chatgpt and gpt-4. *Natural Language Processing Journal*, 6:100048.

Jinliang Lu, Ziliang Pang, Min Xiao, Yaochen Zhu, Rui Xia, and Jiajun Zhang. 2024. Merge, ensemble, and cooperate! a survey on collaborative strategies in the era of large language models. *ArXiv*, abs/2407.06089.

Jing Ma, Chen Chen, Chunyan Hou, and Xiaojie Yuan. 2023. KAPALM: Knowledge grAPh enhAnced language models for fake news detection. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3999–4009, Singapore. Association for Computational Linguistics.

Xiaoxiao Ma, Yuchen Zhang, Kaize Ding, Jian Yang, Jia Wu, and Hao Fan. 2024. On fake news detection with LLM enhanced semantics mining. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 508–521, Miami, Florida, USA. Association for Computational Linguistics.

Mohit Mayank, Shakshi Sharma, and Rajesh Sharma. 2021. Deap-faked: Knowledge graph based approach for fake news detection. *2022 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 47–51.

Qiong Nan, Juan Cao, Yongchun Zhu, Yanyan Wang, and Jintao Li. 2021a. Mdfend: Multi-domain fake news detection. *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*.

Qiong Nan, Juan Cao, Yongchun Zhu, Yanyan Wang, and Jintao Li. 2021b. Mdfend: Multi-domain fake news detection. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 3343–3347.

Qiong Nan, Qiang Sheng, Juan Cao, Beizhe Hu, Danding Wang, and Jintao Li. 2024. Let silence speak: Enhancing fake news detection with generated comments from large language models. *ArXiv*, abs/2405.16631.

Qiong Nan, Qiang Sheng, Juan Cao, Yongchun Zhu, Danding Wang, Guang Yang, Jintao Li, and Kai Shu. 2023. Exploiting user comments for early detection of fake news prior to users' commenting. *ArXiv*, abs/2310.10429.

M. Zohaib Nawaz, M. Saqib Nawaz, Philippe Fournier-Viger, and Yulin He. 2024. Analysis and classification of fake news using sequential pattern mining. *Big Data Mining and Analytics*, 7(3):942–963.

Vy Duong Kim Nguyen and Phuc Do. 2024. Fake news detection using knowledge graph and graph convolutional network. In *Intelligent Systems and Data Science*, pages 216–224, Singapore. Springer Nature Singapore.

F. Olan, U. Jayawickrama, E. O. Arakpogun, J. Suklan, and S. Liu. 2022. Fake news on social media: the impact on society. *Information Systems Frontiers: A Journal of Research and Innovation*, pages 1–16. Advance online publication.

OpenAI. 2022. ChatGPT: Optimizing Language Models for Dialogue. `https://openai.com/blog/chatgpt/`. Accessed: 2023-08-13.

Anjan Pal and Alton Y. K. Chua. 2019. Propagation pattern as a telltale sign of fake news on social media. In *2019 5th International Conference on Information Management (ICIM)*, pages 269–273.

Marco AF Pimentel, Cl'ement Christophe, Tathagata Raha, Prateek Munjal, Praveen K Kanithi, and Shadab Khan. 2024. Beyond metrics: A critical analysis of the variability in large language model evaluation frameworks. *ArXiv*, abs/2407.21072.

Kai Shu, Limeng Cui, Suhang Wang, Dongwon Lee, and Huan Liu. 2019. defend: Explainable fake news detection. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, page 395–405, New York, NY, USA. Association for Computing Machinery.

Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. 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.

Chenguang Song, Yiyang Teng, Yangfu Zhu, Siqi Wei, and Bin Wu. 2022. Dynamic graph neural network for fake news detection. *Neurocomputing*, 505:362–374.

Samia Tasnim, Md Mahbub Hossain, and Hoimonty Mazumder. 2020. Impact of rumors and misinformation on covid-19 in social media. *Journal of preventive medicine and public health = Yebang Uihakhoe chi*, 53(3):171—174.

Kefei Tu, Chen Chen, Chunyan Hou, Jing Yuan, Jundong Li, and Xiaojie Yuan. 2021. Rumor2vec: A rumor detection framework with joint text and propagation structure representation learning. *Information Sciences*, 560:137–151.

Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Neural Information Processing Systems*.

Shubham Vatsal and Harsh Dubey. 2024. A survey of prompt engineering methods in large language models for different nlp tasks. *ArXiv*, abs/2407.12994.

Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science*, 359(6380):1146–1151.

Xin Wang, Yu Zhang, Wenquan Xu, Hanxi Wang, Jingye Cai, Qin Qin, Qin Wang, and Jing Zeng. 2025. Construction of multi-scale fusion attention unified perceptual parsing networks for semantic segmentation of mangrove remote sensing images. *Applied Sciences*, 15(2).

Y. Wang, F. Ma, Z. Jin, Y. Yuan, G. Xun, K. Jha, L. Su, and J. Gao. 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*, pages 849–857.

Yang Wu, Pengwei Zhan, Yunjian Zhang, Liming Wang, and Zhen Xu. 2021. Multimodal fusion with co-attention networks for fake news detection. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2560–2569, Online. Association for Computational Linguistics.

Liang Xiao, Qi Zhang, Chongyang Shi, Shoujin Wang, Usman Naseem, and Liang Hu. 2024. Msynfd: Multi-hop syntax aware fake news detection. *Proceedings of the ACM on Web Conference 2024*.

Xiaoman Xu, Xiangrun Li, Taihang Wang, and Ye Jiang. 2025. Ample: Emotion-aware multimodal fusion prompt learning for fake news detection. In *MultiMedia Modeling*, pages 86–100, Singapore. Springer Nature Singapore.

Yimei Yang, Jinping Liu, Yujun Yang, and Lihui Cen. 2024a. Dual-stream fusion network with multi-head self-attention for multi-modal fake news detection. *Applied Soft Computing*, 167:112358.

Yuzhou Yang, Yangming Zhou, Qichao Ying, Zhenxing Qian, Dan Zeng, and Liang Liu. 2024b. Fact-checking based fake news detection: a review. *ArXiv*, abs/2401.01717.

Litian Zhang, Xiaoming Zhang, Ziyi Zhou, Xi Zhang, Senzhang Wang, Philip S. Yu, and Chaozhuo Li. 2025. Early detection of multimodal fake news via reinforced propagation path generation. *IEEE Transactions on Knowledge and Data Engineering*, 37(2):613–625.

Yongchun Zhu, Qiang Sheng, Juan Cao, Shuokai Li, Danding Wang, and Fuzhen Zhuang. 2022a. Generalizing to the future: Mitigating entity bias in fake news detection. *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Yongchun Zhu, Qiang Sheng, Juan Cao, Qiong Nan, Kai Shu, Minghui Wu, Jindong Wang, and Fuzhen Zhuang. 2022b. Memory-guided multi-view multi-domain fake news detection. *IEEE Transactions on Knowledge and Data Engineering*.

A Appendix

In the technical appendix that follows, we present four crucial prompt words specifically related to the large model section, which play a significant role in optimizing interactions and outputs involving the large model. It consists of four parts: Inquiry, Response, Critique, and Synthesis.

Prompt 1: Inquiry Prompt

System Prompt: You are a professional news analysis assistant.
Context Prompt: Please read and evaluate the following content from social media *[content]* whose authenticity is subject to verification. Ask key questions that will help evaluate its authenticity based on the information provided.

Prompt 2: Response Prompt

System Prompt: You are a professional news analysis assistant.
Context Prompt: *[text]* The above is a piece of news content whose authenticity is uncertain. The question raised by another large model about this news is *[question]*. As an experienced news analyst, provide a clear and concise answer based on the previous questions and the provided news content. Maintain professionalism and objectivity in your response and try to provide specific details that support your conclusion.

Prompt 3: Critique Prompt

System Prompt: You are a professional news analysis assistant.
Context Prompt: *[text]* The above is a piece of news content whose authenticity is uncertain. The question you raised about this news is *[question]*. The answer provided by the other model is *[answer]*. As a rigorous news analyst, raise further questions based on the news content, the questions asked, and the answers given. Your goal is to test the reasonableness and completeness of the existing answers, while identifying any potential logical flaws or inconsistencies. Ensure your questions are constructive and concise.

Prompt 4: Synthesis Prompt

System Prompt: You are a professional news analysis assistant.
Context Prompt: The following is a piece of social media content whose authenticity cannot be confirmed: *[text]*. The question raised by another large model regarding this content is *[question]*. The answer provided by you to this question is *[answer]*. Another big model questions this answer with the query *[query]*. Summarize and analyze the above conversation, integrating all relevant information to form a coherent and logically rigorous analysis. Ensure the summary is concise and to the point, retaining only the information that helps in assessing authenticity.