When Evidence Falls Short: Router-Guided Fake News Detection with Pattern Augmentation

Yujing Wang, **Xiaobao Wang**, **Yiqi Dong**, **Yueheng Sun**, **Di Jin** and **Dongxiao He***
College of Intelligence and Computing, Tianjin University, Tianjin, China
{wangyujing, wangxiaobao, dongyiqi, yhs, jindi, hedongxiao}@tju.edu.cn

Abstract

With the growing complexity of online information, trustworthy fake news detection has become increasingly critical. Although Large Language Models (LLMs) exhibit a strong ability to leverage factual evidence for verification, they remain highly vulnerable to unreliable, noisy, or scarce evidence, undermining robustness in real-world scenarios. Given the generalizability of deceptive patterns in fake news, we consider pattern as a complementary signal under insufficient evidence during factual verification. However, due to LLMs' lack of expertise in deception-specific patterns, realizing such effective collaboration remains challenging. To address these issues, we propose a **R**outer-**G**uided Fake News Detection Framework with **P**attern **A**ugmentation (**RGPA**). Specifically, we introduce a hierarchical routing mechanism including a case router and an external evidence router. It guides news to appropriate reasoning paths adaptively based on a multi-dimensional quality assessment, prioritizing high-quality evidence while mitigating noise. Furthermore, we design an expert model to capture deceptive features and integrate them into LLMs' reasoning, enabling a synergy of factual verification and pattern awareness under evidence-scarce scenarios. Extensive experiments on two real-world datasets demonstrate that RGPA significantly outperforms existing approaches.

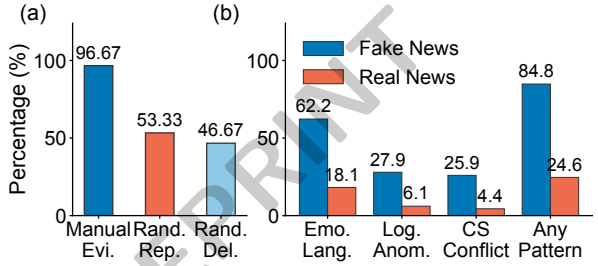

Figure 1: (a) Comparison of LLM-based verification performance on 30 manually selected Weibo samples with sufficient evidence versus randomly perturbed evidence (partial replacement/deletion). (b) Statistics of the three deceptive patterns across the Weibo dataset.

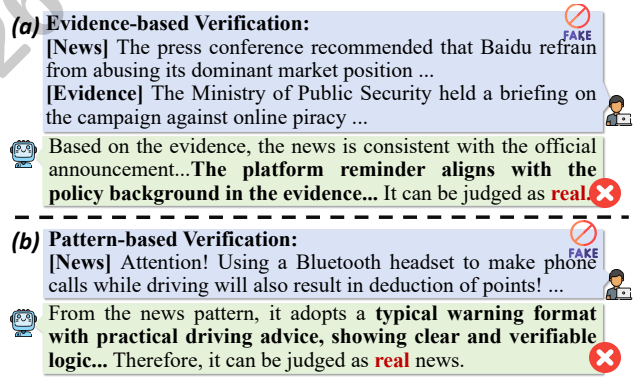

Figure 2: LLMs are susceptible to hallucinatory reasoning when (a) evidence is insufficient and (b) lacks deception-specific patterns.

1 Introduction

With the rapid growth of online platforms, misinformation spreads at unprecedented speed and scale, posing serious threats to public safety [Shu *et al.*, 2020; Liu *et al.*, 2025; Si *et al.*, 2026]. Therefore, it is critical to develop reliable Fake News Detection (FND) systems that can operate under complex environments. Early fake news detection evolved from shallow linguistic features [Feng *et al.*, 2012] to deep pattern recognition [Shu *et al.*, 2020]. However, these methods falter when misinformation is factually grounded rather than stylistically distinct. While evidence-based methods [Popat *et al.*, 2018; Wu *et al.*, 2021] address this by leveraging external knowledge, the traditional model often lacks the reasoning depth and interpretability [Hu *et al.*, 2024]. Recently, Large Language Models (LLMs) have demonstrated strong language understanding abilities and are considered a promising direction. Some works treat LLMs as auxiliary modules that enhance performance [Wang *et al.*, 2025; He *et al.*, 2025a], yet they fail to fully unlock the LLM's inherent reasoning capabilities.

LLMs' strong zero-shot and few-shot generalization enables their effective use in news verification [Pan *et al.*, 2023; Caramancion, 2023; Jin *et al.*, 2025a], often integrated with Retrieval-Augmented Generation (RAG) to directly feed re-

*Corresponding Author

trieved evidence into the prompt context. Indeed, real-world news often exists within a noisy environment where high-quality evidence is not always accessible. To evaluate LLMs' performance under such complexity, we curated samples on the Weibo dataset with high-quality evidence, then randomly removed or replaced evidence partly for LLMs verification. The performance in Figure 1(a) shows that the LLM demonstrates a powerful capacity for fact-based verification when provided with high-quality evidence, but exhibits a sharp drop with degraded evidence, suggesting high sensitivity to noise or scarcity. Figure 2(a) illustrates an example, where the LLM generates hallucinatory reasoning when evidence is insufficient, mistakenly categorizing news as real. While evidence-grounded factual verification is a promising direction, the observed vulnerability reveals a critical limitation of LLM-based detectors: relying solely on retrieved evidence is inadequate for reliable detection in uncertain environments.

In fact, the intrinsic patterns of news content exhibit a certain degree of generalizability [Sheng *et al.*, 2021; Dong *et al.*, 2024] and have long been recognized as discriminative cues in traditional FND research [Ajao *et al.*, 2019; Vaibhav *et al.*, 2019; Dun *et al.*, 2021]. These deceptive patterns can be broadly categorized into three types: Emotional Language, Logical Anomaly, and Commonsense Conflict. Figure 1(b) presents a analysis on the Weibo dataset, quantifying the percentage of each and at least one type in news, revealing substantial discrepancies between fake and real news across three patterns, with these patterns covering the majority of misinformation. This observation inspires us to treat pattern-based features as crucial complementary signals when evidence is insufficient. Therefore, we explore synergizing fact-based verification with pattern awareness within an LLM-based framework to achieve more robust detection under complex evidence conditions. However, realizing such effective collaboration remains challenging, as LLMs lack expertise in implicit patterns due to the absence of task-specific training. As illustrated in Figure 2(b), the LLM fails to infer sensationalistic deceptive intent, leading to incorrect judgments.

Drawing inspiration from intelligent routing techniques in multi-LLM systems [Ong *et al.*, 2024], which dynamically select optimal reasoners, we treat the fake news detection task as an adaptive selection of reasoning paths conditioned on evidence quality in order to achieve effective collaboration. To this end, we propose a **R**outer-**G**uided Fake News Detection Framework with **P**attern **A**ugmentation **(RGPA)**. Specifically, we design a hierarchical dual-routing mechanism to enable evidence perception and utilization dynamically, while integrating pattern insights under insufficient evidence. Leveraging the recurrent nature of rumors [Shin *et al.*, 2018] to prioritize more reliable verified historical cases, we employ cascaded Case and Evidence Routers to evaluate retrieval quality across multiple dimensions and adaptively guide news into three reasoning paths: (i) Case-based Reasoning for high-quality historical matches; (ii) Evidence-based Reasoning when external corroboration is sufficient; and (iii) Integrated Reasoning with Evidence and Pattern-Augmented when retrieval is insufficient. In the third path, to deeply capture the pattern feature inherent in news, we design a Fine-grained Evidence-Pattern Expert that models fragmented evidence-claims via graph-based interactions and learns discriminative attribution scores for the three pattern types: language style, causal logic, and commonsense knowledge through cross-attention mechanisms. By training independently, we extract expert insights to augment LLMs' perception of deception-specific patterns. Importantly, this lightweight design enables low-cost, continuous updates to keep pace with evolving patterns.

In summary, our work makes several key contributions:

- We are the first to explore mitigating LLMs' high sensitivity to fluctuating evidence by incorporating deception patterns into LLM-based fake news detection.
- We propose RGPA, an adaptive framework that dynamically steers reasoning based on evidence quality. By integrating a specialized expert to distill pattern insights, it achieves a synergy between factual verification and pattern awareness within LLM-based detection.
- Extensive experiments on two real-world datasets show that RGPA achieves state-of-the-art results on accuracy, F1, and other key metrics.

2 Methodology

Figure 3 illustrates the overall architecture of RGPA. Firstly, model retrieves historical cases or external evidence related to the news. Then, the hierarchical routers guide the news to different paths based on the quality of the evidence. When evidence is insufficient, we extract fragmented evidence and utilize the pattern insights extracted by the fine-grained expert to enable the deep collaboration of fact verification and pattern perception by the LLM.

2.1 Task Definition

We study the detection of fake news with an LLM $\mathcal{L}$. Given a news dataset $\mathcal{D}$, each instance $x \in \mathcal{D}$ represents a news article. The goal is to predict its veracity label $y \in \{\text{Real}, \text{Fake}\}$. To support reasoning, the news article x is augmented with retrieved evidence or supplementary knowledge S. Formally, the prediction is defined as: $y = \mathcal{L}(x, S)$.

2.2 Evidence Evaluation and Hierarchical Router

External evidence retrieved from open environments can often be unreliable. In contrast, verified historical cases provide a more trustworthy foundation, especially given the recurrent nature of rumors [Shin *et al.*, 2018]. Building on this reliability hierarchy, we design a hierarchical routing mechanism that prioritizes historical consensus. If historical precedents are inconclusive, RGPA then evaluates external evidence, and ultimately resorts to integrated reasoning enhanced by expert patterns when both evidence sources fall short.

Case Evaluator and Router. Given a news item x, we retrieve the top-k semantically similar historical cases from a validated case library D_c. Using a pre-trained language model M, the news and case representations are encoded as:

$$H_x = \text{M}(x), \tag{1}$$

$$H_j = \text{M}(c_j), \quad c_j \in D_c. \tag{2}$$

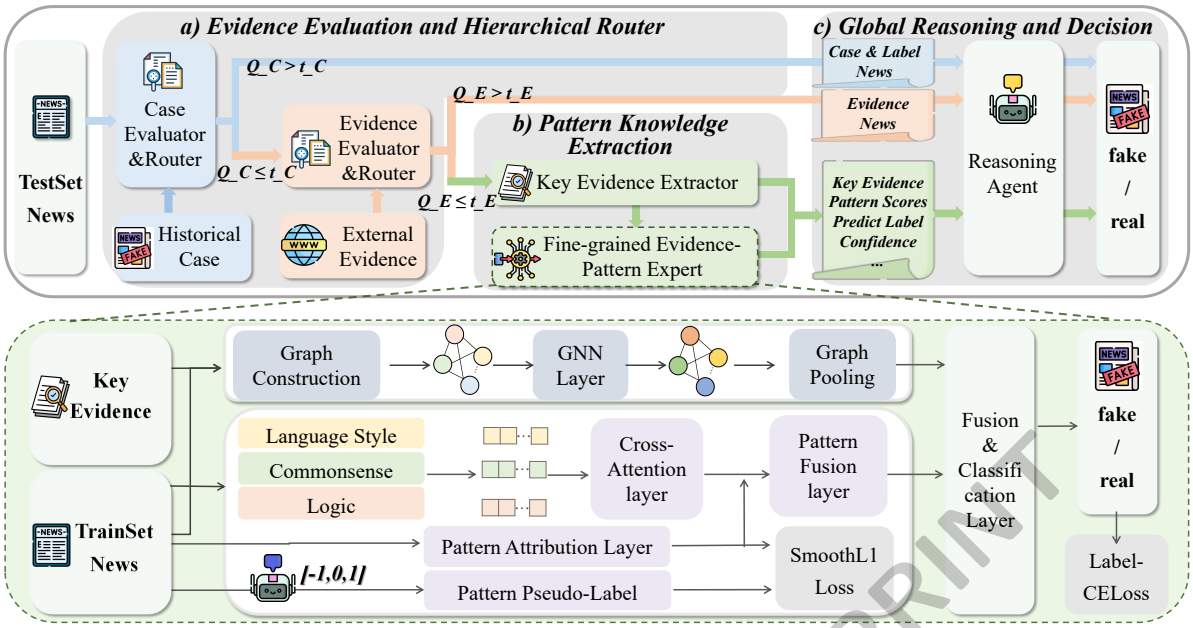

Figure 3: The architecture of our proposed model, which contains three specific steps: (1) Evidence Evaluation and Hierarchical Router: evaluates evidence quality and guides news to different reasoning paths; (2) Pattern Knowledge Extraction: extracts key evidence and specialized pattern knowledge; (3) Global Reasoning and Decision: performs reasoning along each paths to make the final prediction.

We then apply a k-nearest neighbor (kNN) search strategy to retrieve the top-k most similar cases:

$$C_x = \text{TopK}_{c_j \in D_c}\big(\text{sim}(H_x, H_j)\big), \tag{3}$$

where C_x denotes the retrieved case set, $\text{sim}(\cdot,\cdot)$ is a semantic similarity function, and $\text{TopK}(\cdot)$ selects the k cases with the highest similarity scores.

While historical cases are validated, similarity-based retrieval alone is insufficient for reliable judgments. Moreover, direct reliance on LLM-generated confidence is problematic due to the issue of model overconfidence [Chhikara, 2025]. Therefore, we pre-evaluate retrieved cases across multiple dimensions to ensure sufficiency. Specifically, each case set C_x is assessed with prompts p_c with Case Evaluator along three dimensions: $Relevance$, $Consistency$, and $Coverage$. The normalized scores in $[0, 1]$ are averaged to produce an overall case quality score Q_C:

$$Q_C = \text{CaseEvaluator}(p_c, C_x, x). \tag{4}$$

After obtaining the case quality score Q_C, it is compared with a predefined threshold τ_C to trigger first-level routing. If $Q_C > \tau_C$, the historical cases are deemed sufficient, and the news is routed to the case-based reasoning path (Section 2.4), enabling efficient reuse of prior knowledge while reducing evidence noise and inference cost.

External Evidence Evaluator and Router. When historical cases are insufficient ($Q_C \leq \tau_C$), evidence set $E_x = \{e_i\}_{i=1}^{k}$ is retrieved from external sources using the vector similarity retrieval strategy as in Eq.(3). As the credibility of external evidence is often uncertain, we extend the evaluation criteria by introducing a credibility dimension. Each E_x is evaluated with task-specific prompts p_e with Evidence Evaluator across four dimensions: $Relevance$, $Consistency$, $Coverage$, and $Credibility$. Normalized scores are averaged to obtain the overall score Q_E:

$$Q_E = \text{EvidenceEvaluator}(p_e, E_x, x). \tag{5}$$

Compared with predefined external evidence router threshold τ_E, if $Q_E > \tau_E$ the news enters the evidence-based reasoning path (detailed in Section 2.4). Otherwise ($Q_E \leq \tau_E$), we activate the pattern knowledge extraction module, leveraging limited key evidence and deception-specific pattern knowledge to assist the Reasoning Agent.

2.3 Pattern Knowledge Extraction

Key Evidence Extractor. Despite the insufficiency of external evidence, directly discarding it risks the loss of relevant, critical information. Therefore, we employ a Key Evidence Extractor to distill valuable fragments from the external evidence. Simultaneously, we decompose the news x into sub-claims to facilitate a finer-grained alignment between the news and evidence. Guided by task-specific prompts p_k, this process is formulated as:

$$X, K = \text{KeyEvidenceExtractor}(p_k, E_x, x), \tag{6}$$

where $X = \{x_i\}_{i=1}^{n}$ and $K = \{k_j\}_{i=1}^{m}$ represent the sets of decomposed news segments and extracted key fragments.

Fine-Grained Evidence-Pattern Expert. Given the inherent limitations of LLMs in perceiving domain-specific deceptive patterns, we introduce a Fine-grained Evidence–Pattern Expert to extract pattern cues as supplemental knowledge, thereby enabling multi-perspective augmented reasoning. **(1) News-Evidence Graph Interaction.** To

capture the multi-faceted interplay between news and fragments ranging from direct verification to latent complementarity, we construct a fully connected heterogeneous graph $\mathcal{G} = (X \cup K, \mathcal{E})$. Here, the edge set $\mathcal{E} = \{E_{\text{n}}, E_{\text{e}}, E_{\text{n}\leftrightarrow\text{e}}\}$ encodes semantic dependencies: intra-type edges E_{n} and E_{e} capture the internal correlations within news and evidence nodes, respectively, while inter-type edges $E_{\text{n}\leftrightarrow\text{e}}$ model the cross-validation or contradiction between the two node types. Nodes including news fragments X and evidence fragments K are initialized using BERT:

$$H^{(0)} = [\text{BERT}(X); \text{BERT}(K)]. \tag{7}$$

The representations are iteratively refined via an L-layer Heterogeneous Graph Transformer (HGT) [Hu *et al.*, 2020]. For each node v, the update at layer l is defined as:

$$H_v^{(l+1)} = \text{HGT}^{(l)}\Big(H_v^{(l)},\ \{H_u^{(l)} \mid (u,v) \in \mathcal{E}\}\Big), \tag{8}$$

where $\{H_u \mid (v,u) \in \mathcal{E}\}$ represents the information from all adjacent nodes connected to node v, and $H_v^{(l+1)}$ is the embedding of node v after the update in the $(l+1)$-th layer. After L-layers of HGT updates, we obtain the final node representation $H_{\text{out}} \in \mathbb{R}^{(m+n)\times d}$. We then use a pooling layer to obtain the global representation H_e:

$$H_e = \text{Pool}(H_{\text{out}}^{(1)}, \ldots, H_{\text{out}}^{(m+n)}), \tag{9}$$

where Pool represents the average pooling operation.

(2) Fine-grained Pattern Awareness. To capture pattern knowledge, we implement fine-grained awareness across above three generalized perspectives. Firstly, we use an LLM to generate three pattern-related textual descriptions: Language Style A_{el}, Logic A_{cl}, and Commonsense A_{ck}, which are instance-specific interpretations, serving as semantic anchors to guide pattern-aware representation learning. Then we encode the news x and three descriptions using BERT:

$$H_x = \text{BERT}(x), \quad H_{A_p} = \text{BERT}(A_p), \quad p \in \{el, cl, ck\}, \tag{10}$$

where H_x denotes the feature representation of the news item, and H_{A_p} represents the embedding of the pattern description associated with pattern p. To estimate the extent to which a news item exhibits different deception-related patterns, we design a pattern attribution layer that predicts pattern-specific scores based on the news representation:

$$\mathbf{w} = \tanh(\text{MLP}(H_x)), \tag{11}$$

where $\mathbf{w} = [w_{el}, w_{cl}, w_{ck}]^\top$ denotes the signed tendency scores for language style, logic, and commonsense patterns, respectively. The $\tanh(\cdot)$ activation constrains each score to the range $(-1, 1)$, indicating whether a pattern is positively expressed, neutral, or negatively deviated in the news.

To prevent pattern-specific scores from semantic drift or being non-interpretable, we introduce an pseudo-labeling mechanism as auxiliary supervision. Although LLMs lack specialized expertise in deception-specific patterns, their general semantic understanding provides coarse directional signals ensuring pattern scores remain aligned with understandable concepts. Specifically, for each news x, we employ an LLM to assign pseudo-labels to the three pattern channels, indicating whether the corresponding pattern exhibits positive, neutral, or negative:

$$\mathbf{p}_{pl} = [p_{el}, p_{cl}, p_{ck}]^\top, \tag{12}$$

where each element $p_* \in \{-1, 0, 1\}$ serves as a ternary indicator: -1 denotes a pattern deviation or abnormality, 0 indicates irrelevance or absence, and 1 signifies a normal or supportive attribute. To align the predicted attribution scores $\mathbf{w}$ with the LLM-guided priors, we adopt the Smooth L1 loss as the auxiliary objective:

$$\mathcal{L}_{\text{aux}} = \text{SmoothL1}(\mathbf{w}, \mathbf{p}_{pl}). \tag{13}$$

To integrate pattern knowledge into news representations, we employ a cross-attention mechanism to capture semantic dependencies between the news and each pattern description:

$$f_{\text{cross}}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V, \tag{14}$$

$$H_x^{(p)} = f_{\text{cross}}(H_x, H_{A_p}, H_{A_p}), \quad p \in \{el, cl, ck\}, \tag{15}$$

where Q, K, and V denote the query, key, and value matrices, d_k is the feature dimension, and $H_x^{(p)}$ denotes the news representation enhanced by pattern p. The final pattern-aware news representation is obtained by weighted aggregation:

$$H_p = \sum_{p\in\{el,cl,ck\}} w_p\, H_x^{(p)}, \tag{16}$$

where H_p denotes the pattern-aware fused representation of the news. Finally, the evidence-aware representation H_e and the pattern-aware representation H_p are concatenated and fed into the classification layer:

$$\hat{y} = \text{softmax}([H_e; H_p]W_{\text{cls}} + b_{\text{cls}}), \tag{17}$$

where W_{cls} and b_{cls} are learnable parameters. For the fake news detection task, we use the cross-entropy loss function:

$$\mathcal{L}_{\text{cls}} = \text{CrossEntropy}\ (\hat{y}, y). \tag{18}$$

The overall loss of expert is defined as linear combination:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{cls}} + \alpha \cdot \mathcal{L}_{\text{aux}}, \tag{19}$$

where α weights the auxiliary supervision. Despite its specialized proficiency in capturing deceptive patterns through specialized task training, the expert lacks the open-domain knowledge and logical synthesis in LLMs. Consequently, the expert distills discriminative cues and predictive signals to augment the LLM's reasoning process.

2.4 Global Reasoning and Decision

Case-based Reasoning. When the retrieved historical cases satisfy the quality threshold ($Q_C > \tau_C$), the Reasoning Agent executes few-shot reasoning by leveraging the target news x alongside the case-label pairs (C_x, C_y), directly producing the judgment $\hat{y}$ and rationale R, enabling interpretable decision-making.

Evidence-based Reasoning. When high-quality external evidence is available ($Q_C \leq \tau_C$ and $Q_E > \tau_E$), the Agent

Category	Methods (%)	Weibo				Twitter			
		Accuracy	Macro-F1	Precision	Recall	Accuracy	Macro-F1	Precision	Recall
Direct LLM	LLM-zero-shot	59.42	52.87	72.22	59.66	65.08	53.23	66.19	56.63
	LLM-few-shot	61.22	55.29	74.89	61.46	67.63	60.47	67.42	61.15
	LLM-evidence	63.74	61.09	69.26	63.90	67.22	57.19	69.81	59.34
	LLM-evidence-pattern	64.52	61.41	71.95	64.70	67.87	58.27	70.86	60.12
Only-SLM	Bi-LSTM	67.97	67.70	68.69	68.03	78.66	75.81	79.85	74.72
	EANN	71.82	71.56	72.75	71.88	77.68	75.60	77.14	74.91
	BERT-EMO	74.48	73.79	77.62	74.59	78.80	75.97	79.99	74.87
	DeClarE	64.52	64.45	64.59	64.49	76.97	73.95	77.57	73.01
	EVIN	67.58	67.23	68.47	67.65	77.27	74.38	77.80	73.43
	MAC	70.72	70.40	71.76	70.79	80.29	77.86	81.40	76.72
	Pref-FEND	75.27	74.64	78.31	75.37	80.46	79.24	79.55	79.00
	IKA	74.60	74.50	74.60	74.55	81.50	79.40	82.10	78.35
	DM-INTER	68.60	67.34	72.27	68.73	81.86	81.13	80.84	81.58
	Lexi-FEND*	77.30	77.30	77.40	77.40	85.10	84.20	84.45	83.95
Hybrid LLM	Bert+Rationale	73.08	73.06	73.17	73.10	79.01	78.22	77.94	78.71
	SuperICL	70.96	70.13	73.81	71.07	77.48	75.76	76.54	75.29
	ARG	77.15	76.71	79.58	77.24	82.69	81.88	81.73	82.04
	SheepDog	75.43	75.41	75.49	75.42	80.10	77.17	82.39	75.88
	L-Defense	77.71	77.70	77.71	77.71	83.42	82.91	82.54	83.77
	ROE-FND	76.84	76.83	76.93	76.86	84.38	83.11	84.32	82.39
	RGPA(Ours)	**80.77**	**80.67**	**81.49**	**80.82**	**86.14**	**85.34**	**85.51**	**85.19**

Table 1: Experimental results on different datasets. The best performance is in bold. *Resulting numbers are reported by [Yang *et al.*, 2025a]

reasons over the target news x and retrieved evidence set E_x, emphasizing factual grounding to verify claims, predicting label $\hat{y}$ with rationale R.

Integrated Reasoning with Evidence and Pattern-Augmented. When both historical cases and external evidence are insufficient ($Q_C \leq \tau_C$ and $Q_E \leq \tau_E$), direct fact verification is unreliable. To address this, we augment the Reasoning Agent with domain knowledge from the Pattern Expert. The agent receives multi-perspective inputs: target news x, key evidence fragments K_x, pattern attribution scores and the expert's predictions with confidence for both the news and relevant cases. These expert signals convert abstract deceptive patterns into explicit semantic cues, allowing to integrate fragmented evidence with specialized pattern knowledge and perform holistic reasoning to produce the prediction $\hat{y}$ and rationale R. The process is formalized as:

$$\mathcal{C}_x = \{(c_i, y_i, \mathbf{w}_i, \hat{y}_i, conf_i) \mid c_i \in C_x\}, \tag{20}$$

$$\mathcal{T}_x = (x, K_x, \mathbf{w}_x, \hat{y}_x, conf_x), \tag{21}$$

$$\hat{y}, R = \text{IntegratedReasoner}(p_r, \mathcal{C}_x, \mathcal{T}_x), \tag{22}$$

where p_r is a task-specific prompt. In $\mathcal{C}_x$, each tuple contains the case c_i with label y_i, the pattern attribution scores $\mathbf{w}_i$, the predicted label $\hat{y}_i$, and the prediction confidence $conf_i$ from expert. Similarly, $\mathcal{T}_x$ augments the target news x with its fragmented key evidence K_x and the corresponding expert signals $(\mathbf{w}_x, \hat{y}_x, conf_x)$.

3 Experiment

This section presents our experimental evaluation. We first introduce the datasets, baselines, and implementation details (Section 3.1). We then evaluate RGPA on two real-world datasets (Section 3.2), followed by ablation studies (Section 3.3) to assess component contributions. We analyze the routing ratios (Section 3.4) and provide a case study (Section 3.5) to illustrate the effectiveness of hierarchical routing and integrated reasoning.

3.1 Experiment Setting

Dataset. We evaluate on two real-world fake news datasets [Sheng *et al.*, 2021] in Weibo and Twitter. Each contains posts, related external articles, and labels (*Fake/Real*). Weibo includes 6,362 posts, with 3,161 fake and 3,201 real. Twitter consists of 14,709 posts with 5,699 fake and 9,010 real.

Baselines. We compare our model with three groups of baselines: (1) **Direct LLM-based methods:** zero-shot, few-shot with retrieved cases, evidence, and evidence-pattern prompt. We employ Qwen2.5-7B-Instruct[1] for Weibo and Llama-3.1-8B-Instruct[2] for the Twitter dataset as the backbone. The evidence-pattern only instructs the LLM to consider patterns under insufficient evidence, without incorporating any pattern knowledge. (2) **Traditional SLM-based methods:** including Bi-LSTM [Graves and Schmidhuber, 2005], EANN-Text [Wang *et al.*, 2018], BERT-Emo [Zhang *et al.*, 2021], DeClarE [Popat *et al.*, 2018], EVIN [Wu *et al.*, 2021], MAC [Vo and Lee, 2021], Pref-FEND [Sheng *et al.*, 2021], IKA [Guo *et al.*, 2023], DM-INTER [Wang *et al.*, 2024a] , and Lexi-FEND [Yang *et al.*, 2025a]. (3) **Hybrid LLM methods:** including Bert+Rationale, SuperICL [Xu *et al.*, 2024], ARG [Hu *et al.*, 2024], Sheepdog [Wu *et al.*, 2024], L-Defense [Wang *et al.*, 2024b], and ROE-FND [Yang *et al.*, 2025b].

[1]https://huggingface.co/Qwen/Qwen2.5-7B-Instruct

[2]https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct

Implementation Details. We employ Multilingual-E5-Large [Wang *et al.*, 2024c] to generate embeddings for the retrieval of cases and evidence articles. The training set with labels serves as the historical case base, whereas the dataset's articles form the external evidence library. To ensure a fair comparison, both RGPA and all baselines share the same backbone that Qwen2.5-7B-Instruct[3] for Weibo and Llama-3.1-8B-Instruct[4] for Twitter, with a temperature of 0.7. We use GPT-4o generates pseudo-labels for the three patterns. For each news, the Top-5 historical cases and Top-3 external evidence articles are retrieved. Routing thresholds are set via hyperparameter search : case router τ_C = 0.7 (Weibo) / 0.8 (Twitter), evidence router both τ_E = 0.9. Expert models are trained separately for Weibo and Twitter, using bert-base-chinese[5] and bert-base-uncased[6] for text encoding. HGT uses 2 layers, and the auxiliary loss weight α is set to 0.1.

Experimental Setup. GPT inference in RGPA is conducted via OpenAI API [7] , while BERT, Qwen and LLaMA models are locally deployed using HuggingFace Transformers [Wolf *et al.*, 2020]. Experiments are implemented in PyTorch 2.5.0 on four NVIDIA RTX 4090 GPUs.

3.2 Experiment Results

Table 1 shows the performance of the proposed RGPA and baseline models on two datasets. The following conclusions can be drawn: (1) Direct LLM reasoning proves insufficient for fake news detection, achieving only 59.42% accuracy on Weibo and 65.08% on Twitter. Although leveraging retrieved cases or external evidence yields modest gains, LLM-based methods still lag behind traditional detectors due to noisy and incomplete evidence. Even pattern-prompting offers limited improvement, indicating that LLMs lack specialized knowledge of deception-specific patterns and remain prone to hallucinated reasoning. (2) Hybrid frameworks combining LLMs with auxiliary models consistently outperform traditional methods, yet their efficacy remains inconsistent. Approaches like ARG and SheepDog primarily utilize LLMs as auxiliary components without integrating external evidence, which constrains their upper-bound performance. L-Defense and ROE-FND achieve better results by incorporating evidence and multi-step reasoning. However, these methods fail to perceive evidence quality and overlook inherent deception patterns when evidence is sparse, thereby limiting their performance in complex evidence environments. (3) RGPA achieves the best overall performance on both datasets. This advantage stems from two key factors: (i) the hierarchical routing mechanism that adaptively selects reasoning paths based on information quality, prioritizing reliable cases and high-quality evidence to reduce noise; and (ii) the pattern-augmented reasoning strategy that exploits key information and pattern-aware expert knowledge when external evidence is insufficient. These results demonstrate the effectiveness of RGPA for LLM-based fake news detection.

[3]https://huggingface.co/Qwen/Qwen2.5-7B-Instruct
[4]https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct
[5]https://huggingface.co/google-bert/bert-base-chinese
[6]https://huggingface.co/google-bert/bert-base-uncased
[7]https://platform.openai.com

Model	Weibo		Twitter	
	Accuracy	Macro-F1	Accuracy	Macro-F1
w/o Case	79.67	79.66	85.90	85.07
w/o Evidence	73.63	73.60	82.07	81.76
w/o Expert	69.07	68.28	72.15	68.38
w/o CaseR	77.55	77.55	84.34	83.73
w/o EvidenceR	78.96	78.88	85.97	85.15
w/o Pattern-S	75.20	75.19	85.36	84.77
w/o KeyEE	79.67	79.63	84.88	83.91
w/o IntegrR	79.43	79.31	86.01	85.15
RGPA (Ours)	**80.77**	**80.67**	**86.14**	**85.34**

Table 2: Ablation results of RGPA on two datasets (%).

3.3 Ablation Study

Table 2 shows the ablation results on Weibo and Twitter. Removing historical cases (w/o Case) or external evidence (w/o Evidence) hurts performance, especially without evidence, highlighting its importance for LLM reasoning. Crucially, removing all additional insights provided by the experts (w/o Expert) leads to a significant performance drop, indicating that without deception-specific perception, LLMs are prone to hallucination under noisy or scarce evidence. Regarding the routing mechanism, substituting with simple confidence scores (w/o CaseR and w/o EvidenceR) results in performance loss, confirming that LLM suffers from overconfidence [Chhikara, 2025] and fails to reflect true decision reliability. Furthermore, removing the three pattern scores (w/o Pattern-S) and providing only the expert's final prediction to the LLM, or removing the Key Evidence Extractor (w/o KeyEE), leads to a decline in results. This demonstrates the necessity of both explicit pattern features for granular guidance and noise reduction for effective reasoning. Finally, replacing the integration with the expert's raw prediction for the third path (w/o IntegrR) confirms that LLMs effectively synthesize multi-perspective insights better than expert's standalone predictions.

3.4 Routing Analysis

To understand the hierarchical routing mechanism, we visualize the distribution of news across different reasoning paths and their corresponding accuracies. As shown in Figure 5, the results validate the effectiveness of RGPA in dynamic strategy selection. The high accuracy of case- and evidence-based paths confirms that hierarchical routing effectively prioritizes high-quality information to fully leverage the LLM's fact-verification capabilities. Notably, the relatively low proportion of news routed through these two paths reflects the prevalence of noisy or scarce evidence in real-world scenarios, necessitating pre-inference quality assessment. Under weak-evidence conditions, the pattern-enhanced path markedly outperforms direct LLM prompting (Table 1), demonstrating that specialized pattern expertise provides discriminative cues that reduce uncertainty and mitigate hallucinated reasoning. Furthermore, the distributional variance between datasets reveals that rumor recurrence is more pronounced on Weibo, whereas Twitter offers more reliable external evidence, illustrating RGPA's adaptability across platforms.

Path-1
[News] Coca-Cola admitted that its "Minute Maid Orange" contains the banned pesticide carbendazim and has been blacklisted. O WebLink
[Case 1]"Focus Interview" reported Coca-Cola contained carbendazim. **Fake**
[Case Score] Cor 0.95 | Con 0.95 | Cov 0.9 | Conf 0.9

RGPA: It is determined as **Fake**... Both historical cases involve Coca-Cola and carbendazim, and are labeled fake....

Path-2
[News] Suzuki Automobile is negotiating with Changan Automobile and may withdraw from the Chinese market...Changan Suzuki may disappear...
[Case 1]Sony announced its bankruptcy,business giant has fallen.. **Fake**
[Case Score] Cor 0.3 | Con 0.35 | Cov 0.2 | Conf 0.2
[Evidence 1] Suzuki ended its joint venture with Changan...
[Evidence Score] Cre 0.85 | Cor 0.95 | Con 0.9 | Cov 0.85 | Conf 0.9

RGPA: Evidence 1 and 2 show Suzuki terminated joint venture with Changan... judged as **Real.**

Path-3
[News] Attention everyone! The criminal gang has shifted operation base to Xizang. They use people's compassion and take away students by asking for directions. Multiple disappearance have occurred!
[Expert Insights]{Pattern: Emo -0.83| Com -0.14| Logic -0.68| Pred&Conf: Fake 0.76}
[Key Information] 1. The police in Zengcheng announced abducted ... 2. Rumors of abducting students...
[Case 1] Police arrested a child trafficking... Please forward in seconds! **Fake**
[Expert Insights]{Pattern: Emo -0.87 | Com 0.26 | Logic 0.32 | Pred&Conf: Fake 0.63}
[Case 2] This morning, a three-year-old girl was abducted... Contact... Phone... **Real**
[Expert Insights]{Pattern: Emo 0.64 | Com 0.53 | Logic 0.72 | Pred&Conf: Real 0.85}

RGPA: The target news lacks direct evidence. Although Key Information 1 mentions related incidents, it does not support the claim. Expert scores show anomalies in language style (−0.83) and logical structure (−0.68), suggesting non-neutral reporting. The phrase "multiple disappearances" also lacks details and logical support. Thus, consistent with the experts' view, the news is likely fake. Its stylistic and logical deviations resemble fake news cases (Cases 1 and 3), while real Case 2 is more neutral and detailed, with specific contact information. Consequently...judged as **Fake.**

Figure 4: Case studies of RGPA under different routing decisions. We omit the less important content using [...].

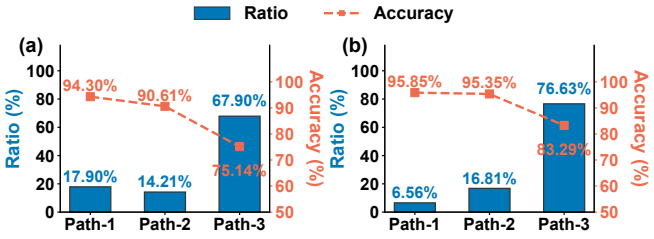

Figure 5: The proportion of routing to the three reasoning paths on the (a) Weibo and (b) Twitter datasets, where path-1 represents the Case-based Reasoning, path-2 represents the Evidence-based Reasoning, and path-3 represents the Integrated Reasoning.

3.5 Case Study

To illustrate the effectiveness of hierarchical routing and expert knowledge augmentation in RGPA, we present representative Weibo cases corresponding to different reasoning paths (Figure 4). In the first case, the claim that "Coca-Cola products contain pesticides" is supported by historical cases that receive high quality scores, allowing to reason directly from historical experience and avoid potential noise introduced by external evidence. In the second case, the claim that "Suzuki exits the Chinese market" is associated with low-quality historical cases, prompting routing to retrieve external evidence, where authoritative reports and official statements provide reliable factual support for the final decision.

In the third scenario, where the retrieval consists only of fragmented reports from other regions without direct factual support, direct LLM reasoning is therefore prone to treating such fragments as related anchors and misclassifying the news. To mitigate this risk, RGPA identifies evidence insufficiency and routes the news to the pattern-augmented reasoning path. Specifically, pattern-specific scores and expert predictions are introduced as supplementary knowledge to capture sensationalized expressions and logical gaps, revealing significant emotional (-0.83) and logical (-0.68) deviations. Guided by these signals, the LLM recognizes rumor-related fragment cues and abnormal pattern deviations, reducing the credibility of the claim. It then reasons by contrasting the pattern–label correspondences across historical cases, observing that the target news closely aligns with fake cases (e.g., Cases 1) in terms of deviation patterns, while sharply diverging from the neutral narration and explicit details of real Case 2, further strengthening the fake judgment. By jointly modeling the multi-perspective of evidence fragment and expert insights, RGPA was able to make reliable judgment.

4 Related Work

Traditional fake news detection methods fall into content-based and evidence-based paradigms. Content-based approaches evolved from handcrafted features [Granik and Mesyura, 2017] to deep models capturing linguistic style and sentiment [Przybyla, 2020; Dong *et al.*, ; Jin *et al.*, 2025b], yet suffer from limited generalization. Evidence-based models employ joint representations of news and evidence[Popat *et al.*, 2018; Wu *et al.*, 2021]. They are constrained by shallow inference and limited interpretability. Recent studies have explored LLMs for fake news detection, either as auxiliary modules providing background knowledge [Hu *et al.*, 2024; He *et al.*, 2025b]. Although they improve accuracy, the auxiliary role of LLMs and reliance on small models constrain overall reasoning capacity. Other studies directly adopt LLMs as direct fact-checkers [Pan *et al.*, 2023]. While these enhance accuracy and interpretability, they remain limited by noisy, incomplete evidence and LLMs' lack of deception-specific knowledge.

5 Conclusion

In this paper, we propose RGPA, a Router-Guided Fake News Detection Framework with Pattern Augmentation that addresses the limitations of LLM-based fake news detection in complex and uncertain environments. By introducing a hierarchical routing mechanism, RGPA prioritizes reliable historical cases and high-quality external evidence to mitigate hallucinated reasoning, while a fine-grained expert module supplements LLMs with deception-specific pattern knowledge when factual evidence is scarce. Extensive experiments on real-world datasets demonstrate the effectiveness of the proposed framework. In future work, we plan to explore continuous learning to adapt to evolving deceptive patterns and develop active investigation agents that enable LLMs to identify information gaps and perform targeted evidence exploration, further enhancing detection robustness.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (No.62276187, 62422210 , 62302333, 92370111, 62272340) .

References

[Ajao *et al.*, 2019] Oluwaseun Ajao, Deepayan Bhowmik, and Shahrzad Zargari. Sentiment aware fake news detection on online social networks. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2507–2511. IEEE, 2019.

[Caramancion, 2023] Kevin Matthe Caramancion. News verifiers showdown: a comparative performance evaluation of chatgpt 3.5, chatgpt 4.0, bing ai, and bard in news fact-checking. In *2023 IEEE Future Networks World Forum (FNWF)*, pages 1–6. IEEE, 2023.

[Chhikara, 2025] Prateek Chhikara. Mind the confidence gap: Overconfidence, calibration, and distractor effects in large language models. *arXiv preprint arXiv:2502.11028*, 2025.

[Dong *et al.*,] Yiqi Dong, Dongxiao He, Xiaobao Wang, Yawen Li, and Xiaowen Su. A generalized deep markov random fields framework for fake news detection.

[Dong *et al.*, 2024] Yiqi Dong, Dongxiao He, Xiaobao Wang, Youzhu Jin, Meng Ge, Carl Yang, and Di Jin. Unveiling implicit deceptive patterns in multi-modal fake news via neuro-symbolic reasoning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 8354–8362, 2024.

[Dun *et al.*, 2021] Yaqian Dun, Kefei Tu, Chen Chen, Chunyan Hou, and Xiaojie Yuan. Kan: Knowledge-aware attention network for fake news detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 81–89, 2021.

[Feng *et al.*, 2012] Song Feng, Ritwik Banerjee, and Yejin Choi. Syntactic stylometry for deception detection. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, pages 171–175, 2012.

[Granik and Mesyura, 2017] Mykhailo Granik and Volodymyr Mesyura. Fake news detection using naive bayes classifier. In *2017 IEEE first Ukraine conference on electrical and computer engineering (UKRCON)*, pages 900–903. IEEE, 2017.

[Graves and Schmidhuber, 2005] Alex Graves and Jürgen Schmidhuber. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural networks*, 18(5-6):602–610, 2005.

[Guo *et al.*, 2023] Hao Guo, Weixin Zeng, Jiuyang Tang, and Xiang Zhao. Interpretable fake news detection with graph evidence. In *Proceedings of the 32nd ACM international conference on information and knowledge management*, pages 659–668, 2023.

[He *et al.*, 2025a] Dongxiao He, Yiqi Dong, Xiaobao Wang, Meng Ge, Carl Yang, Di Jin, and Witold Pedrycz. Unveiling fine-grained deceptive patterns in multi-modal fake news: An explainable neuro-symbolic framework with lvlms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[He *et al.*, 2025b] Jing He, Han Zhang, Yuanhui Xiao, Wei Guo, Shaowen Yao, and Renyang Liu. Factguard: Event-centric and commonsense-guided fake news detection. *arXiv preprint arXiv:2511.10281*, 2025.

[Hu *et al.*, 2020] Ziniu Hu, Yuxiao Dong, Kuansan Wang, and Yizhou Sun. Heterogeneous graph transformer. In *Proceedings of the web conference 2020*, pages 2704–2710, 2020.

[Hu *et al.*, 2024] Beizhe Hu, Qiang Sheng, Juan Cao, Yuhui Shi, Yang Li, Danding Wang, and Peng Qi. Bad actor, good advisor: Exploring the role of large language models in fake news detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 22105–22113, 2024.

[Jin *et al.*, 2025a] Di Jin, Jun Yang, Xiaobao Wang, Junwei Zhang, Shuqi Li, and Dongxiao He. A dynamic knowledge update-driven model with large language models for fake news detection. *arXiv preprint arXiv:2509.11687*, 2025.

[Jin *et al.*, 2025b] Di Jin, Yujun Zhang, Bingdao Feng, Xiaobao Wang, Dongxiao He, and Zhen Wang. Backdoor attack on propagation-based rumor detectors. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 17680–17688, 2025.

[Liu *et al.*, 2025] Zhiwei Liu, Xin Zhang, Kailai Yang, Qianqian Xie, Jimin Huang, and Sophia Ananiadou. Fmdllama: Financial misinformation detection based on large language models. In *Companion Proceedings of the ACM on Web Conference 2025*, pages 1153–1157, 2025.

[Ong *et al.*, 2024] Isaac Ong, Amjad Almahairi, Vincent Wu, Wei-Lin Chiang, Tianhao Wu, Joseph E Gonzalez, M Waleed Kadous, and Ion Stoica. Routellm: Learning to route llms with preference data. *arXiv preprint arXiv:2406.18665*, 2024.

[Pan *et al.*, 2023] Liangming Pan, Xiaobao Wu, Xinyuan Lu, Luu Anh Tuan, William Yang Wang, Min-Yen Kan, and Preslav Nakov. Fact-checking complex claims with program-guided reasoning. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 6981–7004, 2023.

[Popat *et al.*, 2018] Kashyap Popat, Subhabrata Mukherjee, Andrew Yates, and Gerhard Weikum. Declare: Debunking fake news and false claims using evidence-aware deep learning. *arXiv preprint arXiv:1809.06416*, 2018.

[Przybyla, 2020] Piotr Przybyla. Capturing the style of fake news. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 490–497, 2020.

[Sheng *et al.*, 2021] Qiang Sheng, Xueyao Zhang, Juan Cao, and Lei Zhong. Integrating pattern-and fact-based fake

news detection via model preference learning. In *Proceedings of the 30th ACM international conference on information & knowledge management*, pages 1640–1650, 2021.

[Shin *et al.*, 2018] Jieun Shin, Lian Jian, Kevin Driscoll, and François Bar. The diffusion of misinformation on social media: Temporal pattern, message, and source. *Computers in Human Behavior*, 83:278–287, 2018.

[Shu *et al.*, 2020] Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. 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, 2020.

[Si *et al.*, 2026] Aojie Si, Shizhan Chen, Xiaobao Wang, Yueheng Sun, and Jiaai Guo. Beyond semantics: Exploiting propagation structures with dual-adapter llms for fake news detection. *Neural Networks*, page 108904, 2026.

[Vaibhav *et al.*, 2019] Vaibhav Vaibhav, Raghuram Mandyam, and Eduard Hovy. Do sentence interactions matter? leveraging sentence level representations for fake news classification. In *Proceedings of the thirteenth workshop on graph-based methods for natural language processing (TextGraphs-13)*, pages 134–139, 2019.

[Vo and Lee, 2021] Nguyen Vo and Kyumin Lee. Hierarchical multi-head attentive network for evidence-aware fake news detection. *arXiv preprint arXiv:2102.02680*, 2021.

[Wang *et al.*, 2018] Yaqing Wang, Fenglong Ma, Zhiwei Jin, Ye Yuan, Guangxu Xun, Kishlay Jha, Lu Su, and Jing Gao. 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, 2018.

[Wang *et al.*, 2024a] Bing Wang, Ximing Li, Changchun Li, Bo Fu, Songwen Pei, and Shengsheng Wang. Why misinformation is created? detecting them by integrating intent features. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 2304–2314, 2024.

[Wang *et al.*, 2024b] Bo Wang, Jing Ma, Hongzhan Lin, Zhiwei Yang, Ruichao Yang, Yuan Tian, and Yi Chang. Explainable fake news detection with large language model via defense among competing wisdom. In *Proceedings of the ACM Web Conference 2024*, pages 2452–2463, 2024.

[Wang *et al.*, 2024c] Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. Multilingual e5 text embeddings: A technical report. *arXiv preprint arXiv:2402.05672*, 2024.

[Wang *et al.*, 2025] Bing Wang, Ximing Li, Changchun Li, Bingrui Zhao, Bo Fu, Renchu Guan, and Shengsheng Wang. Robust misinformation detection by visiting potential commonsense conflict. *arXiv preprint arXiv:2504.21604*, 2025.

[Wolf *et al.*, 2020] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, et al. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45, 2020.

[Wu *et al.*, 2021] Lianwei Wu, Yuan Rao, Xiong Yang, Wanzhen Wang, and Ambreen Nazir. Evidence-aware hierarchical interactive attention networks for explainable claim verification. In *Proceedings of the twenty-ninth international conference on international joint conferences on artificial intelligence*, pages 1388–1394, 2021.

[Wu *et al.*, 2024] Jiaying Wu, Jiafeng Guo, and Bryan Hooi. Fake news in sheep's clothing: Robust fake news detection against llm-empowered style attacks. In *Proceedings of the 30th ACM SIGKDD conference on knowledge discovery and data mining*, pages 3367–3378, 2024.

[Xu *et al.*, 2024] Canwen Xu, Yichong Xu, Shuohang Wang, Yang Liu, Chenguang Zhu, and Julian McAuley. Small models are valuable plug-ins for large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 283–294, 2024.

[Yang *et al.*, 2025a] Sijia Yang, Xianyong Li, Yajun Du, Dong Huang, Xiaoliang Chen, Yongquan Fan, and Shumin Wang. A hierarchical dual-view model for fake news detection guided by discriminative lexicons. *International Journal of Machine Learning and Cybernetics*, 16(2):1071–1090, 2025.

[Yang *et al.*, 2025b] Yuzhou Yang, Yangming Zhou, Zhiying Zhu, Zhenxing Qian, Xinpeng Zhang, and Sheng Li. Roefnd: A case-based reasoning approach with dual verification for fake news detection via llms. *arXiv preprint arXiv:2506.11078*, 2025.

[Zhang *et al.*, 2021] Xueyao Zhang, Juan Cao, Xirong Li, Qiang Sheng, Lei Zhong, and Kai Shu. Mining dual emotion for fake news detection. In *Proceedings of the web conference 2021*, pages 3465–3476, 2021.