PHPFND: Detecting Fake News via Post-hoc Processing of LLMs Hallucination

Jinke Ma[1], Jiachen Ma[1, 2], Wei Zhang[1*], Yong Liu[1*]

[1]School of Computer Science and Big Data, Heilongjiang University, Harbin, China
[2]School of Artificial Intelligence and Computer Science, Shaanxi Normal University, Xi'an, China
2231968@s.hlju.edu.cn, {majiachen,zhangwei_jsj,liuyong123456}@hlju.edu.cn

Abstract

Large Language Models (LLMs) perform excellently in fake news detection tasks, but their outputs are often accompanied by hallucinations, i.e., generated content that is contradictory to facts. Previous studies have mostly mitigated hallucinations through prompt design. However, this paper reveals that regions in news articles which easily induce hallucinations in LLMs correspond closely to the most challenging regions for fake news detectors. In this paper, we propose a fake news detection framework (PHPFND) based on post-hoc processing of LLMs hallucination. Specifically, our framework includes a hallucination detection module (ISHD) based on information structuring that detects three types of hallucinations in LLMs in a targeted manner, and a hallucination-driven feature enhancement mechanism (HDFE) that incorporates hallucination signals as explicit features into sentence-level encoding and feature fusion to guide the model's attention toward high-risk regions. Experimental results on two mainstream fake news datasets show that our proposed method significantly outperforms LLM-based baselines.

Introduction

Fake news detection, as a key task in the field of natural language processing, has attracted widespread attention in recent years. LLMs such as GPT (Radford et al. 2018) and LLaMA (Touvron et al. 2023), with their powerful language understanding and generation capabilities, have made significant progress in fake news detection tasks (Wang et al. 2024a; Qi et al. 2024; Hu et al. 2024; Wan et al. 2024). However, due to the problem of knowledge hallucination (Ji et al. 2023), LLMs still face inevitable limitations in detecting fake news. Existing studies mainly alleviate hallucination through methods such as Retrieval-Augmented Generation (RAG) (Boumber et al. 2024) and knowledge graphs (Guo et al. 2023), but rarely explore the positive role of hallucination signals themselves in fake news detection. For example, how do the regions of news articles that easily trigger LLMs hallucination differ from other regions? Can we regard these regions as the most challenging cases for fake news detectors, thereby improving their performance?

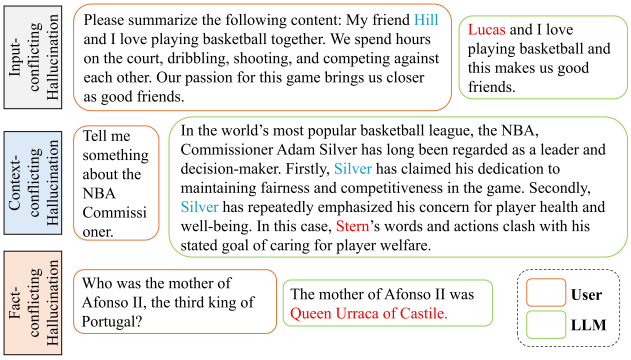

Figure 1: Illustrations of the three types of LLM hallucinations that we define. For input-conflicting hallucination, the LLM makes a mistake in the person's name (*Hill→Lucas*) during summarization. For the context-conflicting hallucination, the LLM discusses *Silver* in the early stage, but later switches to *Stern*, resulting in a contradiction. For the fact-conflicting hallucination, the LLM states the mother of Afonso II was *Queen Urraca of Castile*, while the correct answer is *Dulce Berenguer of Barcelona*

.

Generally speaking, knowledge hallucination refers to the phenomenon where LLMs generate false or misleading information (Ji et al. 2023). Previous studies have summarized the types of LLMs hallucination (Zhang et al. 2025), including input-conflicting, context-conflicting, and fact-conflicting. Input-conflicting hallucination refers to inconsistencies between the model output and user input; context-conflicting hallucination refers to contradictions within the output content; fact-conflicting hallucination refers to model outputs that are inconsistent with world knowledge. Figure 1 shows typical examples of the three types of hallucinations in LLMs.

To alleviate hallucination in a targeted manner, this paper proposes a hallucination detection module (ISHD) based on information structuring, which detects and then corrects the three types of hallucinations. Based on this module, we further find through experiments that the regions in news articles which easily induce hallucination in LLMs correspond closely to the most challenging regions for fake news de-

*Corresponding authors.

tectors. In this paper, we propose a novel fake news detection framework (PHPFND) based on post-hoc processing of LLMs hallucination. This framework first detects hallucination segments through information structuring. It then leverages the detection results to drive multi-round correction of the LLM outputs. Finally, it incorporates the hallucination detection results as explicit features into sentence-level encoding and feature fusion. The main contributions of this paper are as follows:

- We explore the coupling between hallucinations in LLMs and the inherent challenges for fake news detectors, offering a novel perspective on model blind spots.
- We propose a targeted hallucination detection and correction module (ISHD), and further design a feature enhancement mechanism (HDFE) guided by hallucination detection feedback. Their combination improves detection accuracy and explainability.
- Experiments on two mainstream fake news datasets demonstrate the effectiveness and explainability of our proposed method.

Related Work

With the rapid evolution of online social platforms, fake news detection has become a long-term and highly active research area (Ma et al. 2022, 2023, 2024; Dai et al. 2025; Ma et al. 2025). In recent years, LLMs have shown remarkable capabilities in fake news detection, spurring numerous studies. Approaches include generating multiple veracity reasoning paths with debate-like processes (Wang et al. 2024a); leveraging external tools and retrieval to model cues from text, images, and claims for context-free detection (Qi et al. 2024); and revealing LLMs' strengths in content analysis but weaknesses in veracity judgment due to factuality and hallucination challenges (Hu et al. 2024). Many works integrate LLMs as auxiliary feature generators for SLMs or directly apply LLMs with prompting or training strategies (Koike, Kaneko, and Okazaki 2024; Kumarage et al. 2023; Zhang and Gao 2023; Wang, Chang, and Peng 2024; Yang et al. 2024; Zeng and Gao 2023). Hallucination remains an unavoidable issue, attracting extensive research on its causes, classification, and mitigation (Huang et al. 2025), including adversarial evaluations (Lin, Hilton, and Evans 2021), self-reflection (Madaan et al. 2023), contrastive prompting (Zeng and Gao 2023), fact-checking in chain-of-thought reasoning (Dhuliawala et al. 2023), and specialized benchmarks (Guan et al. 2024). Methods to reduce hallucinations include balanced fine-tuning with positive and negative prompts (Liu et al. 2023), iterative concise answer generation (Wang et al. 2024b), and training-free detection and correction (Yin et al. 2024). However, eliminating hallucinations remains fundamentally infeasible under current LLM architectures (Gao et al. 2023; Rafailov et al. 2023; Shinn et al. 2023; Tay et al. 2022), posing a long-term challenge. Unlike prior work, this paper systematically uncovers the intrinsic link between hallucination-triggering input regions and fake news detectors' discrimination difficulties, leveraging this to enhance performance.

Methodology

An overview of our method can be seen in Figure 2.

Hallucination Detection Based on Information Structuring

We extract sentence-level events from the news article $X = \{x_1, \ldots, x_n\}$ and the LLM output $T = \{t_1, \ldots, t_m\}$, where each sentence x_i or t_j is converted into a structured event tuple $e = \{s, v, o, c, l\}$ denoting subject, predicate, object, time, and location. This is achieved using an LLM.

Input-Conflicting Hallucination Detection To detect input-conflicting hallucination, we align events e_i^X and e_j^T extracted from x_i and t_j and compare their slot-level semantics.

Event Alignment. We compute the similarity between each event pair (e_i^X, e_j^T) using a weighted sum of cosine similarities across corresponding slots:

$$sim(e_i^X, e_j^T) = \frac{1}{|S|} \sum_{r \in S} \alpha_r \cdot cos(f_{i,r}^X, f_{j,r}^T), \tag{1}$$

where $S = \{s, v, o, c, l\}$ is the set of slot indices, and α_r is the weight of slot r. $f_{i,r}^X$ and $f_{j,r}^T$ are semantic vectors of slot r in e_i^X and e_j^T, encoded by a pretrained encoder.

To determine the weights α_r for each slot, we use an unsupervised statistical estimation method. Specifically, for a large number of event pairs, we compute the variance σ_r^2 of the semantic similarity distribution of each slot, and then scale the inverse of the variance to obtain the weights:

$$\alpha_r = \frac{\sum_{p \in S}(\sigma_p^2 + \epsilon)}{\sigma_r^2 + \epsilon}, \tag{2}$$

where ϵ is a small smoothing constant. Based on the above event similarity, we construct a similarity matrix $A \in \mathbb{R}^{m \times n}$, where m and n denote the number of events in the news X and the LLM output T, respectively. When $m \neq n$, we introduce $|m-n|$ virtual events to expand A into a square matrix $A' \in \mathbb{R}^{k \times k}$ with $k = \max(m, n)$. The similarity between any virtual event and real event is set to a constant $\delta = -1$, enforcing minimal similarity to avoid unintended matches:

$$A'_{i,j} = \begin{cases} sim(e_i^X, e_j^T), & \text{if } i \leq m,\ j \leq n \\ \delta = -1, & \text{otherwise} \end{cases}. \tag{3}$$

We then apply the Hungarian algorithm to A' to solve the maximum weight matching problem and obtain the optimal alignment $\mathbf{G}$ between e^X and e^T:

$$\mathbf{G} = \arg\max_{g \in \Pi} \sum_{i=1}^{k} A'_{i,g(i)}, \tag{4}$$

where Π denotes the set of all valid one-to-one mappings, and $g(i)$ is the index of the output event e_j^T matched to the i-th news event e_i^X. If $g(i)$ points to a virtual event, e_i^X is unmatched and excluded from subsequent comparisons. Conversely, when $m < n$, we perform matching from T to X using $g(j)$ analogously. To eliminate spurious matches, we

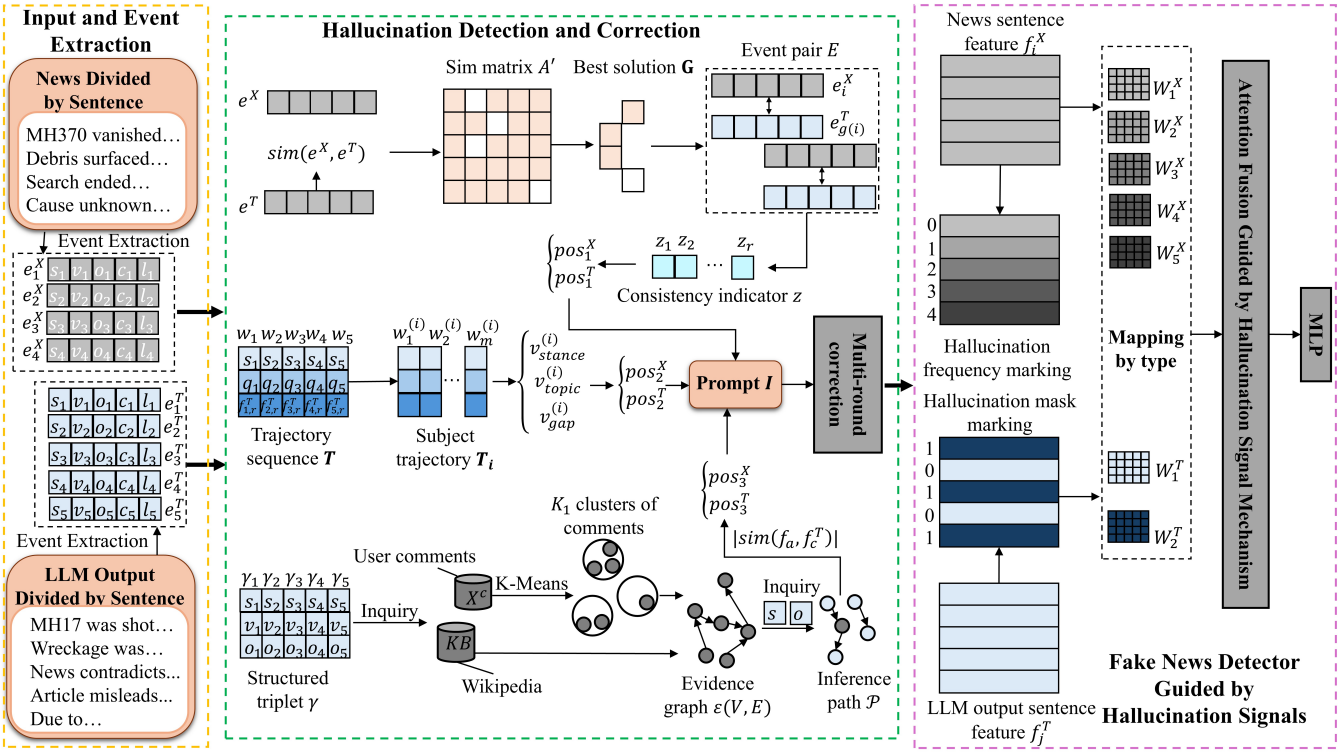

Figure 2: An Overview of the PHPFND.

impose a threshold ϕ_1 and retain only aligned pairs where $sim(e_i^X, e_{g(i)}^T) > \phi_1$.

Event Comparison. For each matched pair $E_i = \{e_i^X, e_{g(i)}^T\}$, we assess slot-level consistency. Given a similarity threshold τ for all slots and using the similarities stored in the matrix A', we define the slot-consistency indicator function z_r as:

$$z_r = \begin{cases} 0, & \text{if } cos(f_{i,r}^X, f_{g(i),r}^T) \geq \tau \\ 1, & \text{otherwise} \end{cases}. \tag{5}$$

A hallucination is detected if:

- the subject s or the object o does not match (but not both);
- s and o match, but v, c, or l does not match.

where 0 indicates consistency and 1 indicates inconsistency. Missing slots are marked as consistent. We record positions of input-conflicting hallucination in X and T as pos_1^X and pos_1^T, respectively.

Context-Conflicting Hallucination Detection To detect context-conflicting hallucination, we propose EST (Entity-Stance Trajectory) modeling, which traces the evolution of subject entities across sentences to identify contextual inconsistencies.

Subject Trajectory Modeling. Given the LLM output $T = \{t_1, \ldots, t_m\}$, we first perform coreference resolution by an LLM to identify the subject s_j in each sentence t_j. Then, we compute the stance polarity $q_j \in [-1, 1]$ of each sentence t_j toward its subject s_j using a pretrained encoder. A higher q_j (closer to 1) suggests a supportive stance, while a lower q_j (closer to -1) indicates an opposing stance. We also extract the semantic embedding of each sentence, denoted as f_j^T. Each sentence t_j is then represented as a node $w_j = \{s_j, q_j, f_j^T\}$, and the output sequence forms the initial trajectory $\mathcal{T} = [w_1, w_2, \ldots, w_m]$.

We group nodes sharing the same subject s_i into a trajectory $\mathcal{T}_i = [w_1^{(i)}, w_2^{(i)}, \ldots, w_{|\mathcal{C}_i|}^{(i)}]$, indexed by $\mathcal{C}_i = \{j \mid s_j = s_i\}$. Trajectories where s_i occurs only once are discarded.

Trajectory Consistency Scoring. To evaluate the contextual consistency of each trajectory $\mathcal{T}_i$, we compute three indicators:

Stance Variation. This indicator measures the fluctuation in stance polarity $q_j^{(i)}$ between adjacent sentences with the same subject s_i:

$$V_{\text{stance}}^{(i)} = \frac{1}{|\mathcal{C}_i| - 1} \sum_{j=2}^{|\mathcal{C}_i|} \left| q_j^{(i)} - q_{j-1}^{(i)} \right|. \tag{6}$$

Higher values indicate abrupt attitude shifts toward the subject, suggesting inconsistency.

Semantic Variation. This indicator measures semantic drift in the trajectory by calculating the cosine similarity between the embeddings $f_j^{T,(i)}$ of sentences with the same topic s_i:

$$V_{\text{topic}}^{(i)} = \frac{1}{|\mathcal{C}_i| - 1} \sum_{j=2}^{|\mathcal{C}_i|} \left(1 - cos(f_j^{T,(i)}, f_{j-1}^{T,(i)})\right). \tag{7}$$

Larger values indicate low coherence and potential topic inconsistency.

Coherence Gap. Let $\{j_1^{(i)}, \ldots, j_{|\mathcal{C}_i|}^{(i)}\}$ be the global indices of trajectory nodes. The coherence gap is defined as:

$$V_{\text{Gap}}^{(i)} = \max_{1 \le k < |\mathcal{C}_i|} \left(j_{k+1}^{(i)} - j_k^{(i)} - 1 \right). \tag{8}$$

It captures the longest span where the subject s_i is omitted in consecutive sentences. A larger gap implies worse contextual continuity.

Each metric captures a distinct type of contextual instability. If any of the three scores exceeds its corresponding threshold ϕ_2, ϕ_3, or ϕ_4, the trajectory is flagged as hallucinated. The detected context-conflicting hallucination positions in X and T are recorded as pos_2^X and pos_2^T, respectively.

Fact-Conflicting Hallucination Detection To detect fact-conflicting hallucination, we combine multi-source evidence retrieval and semantic path inference.

Multi-Source Evidence Retrieval. For each sentence t_j in the LLM output T, we extract its subject-verb-object triple $\gamma_j = \{s_j, v_j, o_j\}$ as a claim. Two types of evidence are retrieved for γ_j: 1) factual knowledge from Wikipedia KB; 2) user comments X^c associated with the news article X.

For evidence type 2), we semantically encode all comments, apply K-Means clustering, and extract K_1 representative cluster centroids as evidence. All evidence is preprocessed with LLM-based coreference resolution and represented in triple form:

$$\gamma_i^{(j)} = \left(s_i^{(j)}, v_i^{(j)}, o_i^{(j)} \right). \tag{9}$$

Semantic Path Reasoning. We build a hybrid evidence graph $\mathcal{E}_j = (V_j, E_j)$ for each γ_j using retrieved triples $\gamma_i^{(j)}$. The node set comprises all subjects and objects:

$$V_j = \left\{ s_i^{(j)},\ o_i^{(j)} \mid 1 \le i \le n \right\}. \tag{10}$$

The edge set E_j consists of directed edges, each representing a predicate $v_i^{(j)}$ linking the subject $s_i^{(j)}$ to the object $o_i^{(j)}$:

$$E_j = \left\{ \left(s_i^{(j)}, v_i^{(j)}, o_i^{(j)} \right) \mid 1 \le i \le n \right\}. \tag{11}$$

Starting from the claim subject s_j, we traverse the evidence graph $\mathcal{E}_j$ to find all semantic paths leading to the object o_j. Let the path set be:

$$\mathcal{P}_j = \{a \mid a : s_j \rightsquigarrow o_j,\ a \in \mathcal{E}_j\}. \tag{12}$$

For each path $a \in \mathcal{P}_j$, we obtain an abstracted predicate $v_j^{s,o}$ representing the inferred relation between s_j and o_j, defined as:

$$v_j^{s,o} = \begin{cases} v_i^{(j)}, & single-hop \\ \mathcal{R}(a), & multi-hop \end{cases}. \tag{13}$$

Here, $\mathcal{R}(\cdot)$ denotes a reasoning abstraction function implemented by an LLM to summarize multi-hop paths into a condensed relation.

Since the claim predicate v_j may differ from the evidence relation $v_j^{s,o}$, we evaluate their consistency by computing the cosine similarity between their semantic vectors: f_a (from the evidence path) and f_c^T (from the claim),

$$sim(v_j, v_j^{s,o}) = cos(f_a,\ f_c^T). \tag{14}$$

We rank paths by similarity and select the top K_2 triples from them. Based on the ratio of contradictory to supportive evidence among them, we determine whether the claim γ_j is hallucinated.

We finally obtain the position indices of fact-conflicting hallucination pos_3^T and pos_3^X for the LLM output and corresponding news sentences.

Hallucination Correction To perform targeted correction, we design a dynamic prompting mechanism tightly integrated with the detection module. Let $pos = \{pos_1, pos_2, pos_3\}$ and $h = \{h_1, h_2, h_3\}$ denote the hallucination positions and types, respectively. These are inserted into a predefined template I to form a correction-oriented prompt $p = \{I, X, pos, h\}$. The template I includes three variants corresponding to different hallucination types: $I = \{I_1, I_2, I_3\}$. An example of template I can be found in Section 6.

Since a single revision often fails to fully resolve contradictions, we introduce an iterative correction process. In each round i, the hallucination detector yields $H_i = \{pos, h\}$, which guides the LLM to revise the output. Iteration continues until no hallucination is detected or the maximum number of steps K_3 is reached. Our approach differs from existing methods in its explicit use of pos and hallucination-type-specific prompts, enabling more precise correction and reducing the risk of generalization failure caused by uniform prompting.

Fake News Detector Guided by Hallucination Signals

After multiple rounds of correction, we integrate the corrected LLM outputs $T = \{t_1, \ldots, t_m\}$, the hallucination-indicator signals H from the hallucination detector, and the news sentences $X = \{x_1, \ldots, x_n\}$. To leverage these signals for more reliable fake news detection, we propose a post-hoc hallucination–guided framework (PHPFND) consisting of two parts: a Hallucination-Driven Feature Enhancement (HDFE) mechanism and a Classification and Loss module.

Hallucination-Driven Feature Enhancement Mechanism The feature enhancement process comprises three components: 1) hallucination masking, which uses the indicator from the final detection round to suppress residual hallucination noise; 2) feature separation, which leverages per-round hallucination signals to highlight high-risk semantics; and 3) attention fusion, which aggregates processed features with weights guided by hallucination signals.

Hallucination Masking. Even after multiple correction rounds, residual hallucinations may remain. For each generated sentence t_j, we construct a hallucination-indicator vector $H_j^T = (h_1^{(j)}, h_2^{(j)}, h_3^{(j)})$ covering three hallucination types and generate a binary mask sequence $M = \{m_1, \ldots, m_m\}$, where $m_j = 1$ indicates the presence of a hallucination.

Feature Separation. Each news sentence x_i and each generated sentence t_j are encoded as $f_i^X \in \mathbb{R}^d$ and $f_j^T \in \mathbb{R}^d$ using a pretrained encoder. During multi-round correction, the hallucination frequency of x_i, denoted as $u_i \in \{0, 1, \ldots, K_3 + 1\}$, is recorded. Based on u_i, x_i is assigned to one of K_3+2 reliability categories, each with a projection matrix $W_k^X \in \mathbb{R}^{d \times d'}$:

$$\hat{f}_i^X = (W_{u_i}^X)^\top f_i^X. \tag{15}$$

For generated sentences t_j, two projection matrices W_1^T and W_2^T are used for hallucination-free ($m_j = 0$) and hallucinated ($m_j = 1$) cases, respectively:

$$\hat{f}_j^T = \begin{cases} (W_1^T)^\top f_j^T, & m_j = 0 \\ (W_2^T)^\top f_j^T, & m_j = 1 \end{cases}. \tag{16}$$

Attention Fusion Guided by Hallucination Signals. Hallucination signals play two roles: (1) they alter feature subspaces via branch selection; and (2) they act as additive priors in the following attention module, amplifying high-u_i news sentences and suppressing hallucinated ones (m_j=1).

Attention scores are computed as:

$$\alpha_i^X = \frac{\exp\Big((f_q^X)^\top \hat{f}_i^X + \beta_1 u_i\Big)}{\sum_{k=1}^n \exp\Big((f_q^X)^\top \hat{f}_k^X + \beta_1 u_k\Big)}, \tag{17}$$

$$\alpha_j^T = \frac{\exp\Big((f_q^T)^\top \hat{f}_j^T - \beta_2 m_j\Big)}{\sum_{k=1}^m \exp\Big((f_q^T)^\top \hat{f}_k^T - \beta_2 m_k\Big)}, \tag{18}$$

where f_q^X and f_q^T are learnable query vectors, and β_1 and β_2 are trainable scalars that control the signal strength.

The weighted sentence representations are then aggregated as:

$$\mathbf{v}^X = \sum_{i=1}^n \alpha_i^X \hat{f}_i^X, \quad \mathbf{v}^T = \sum_{j=1}^m \alpha_j^T \hat{f}_j^T. \tag{19}$$

Finally, a gated fusion module generates the global representation:

$$\mathbf{v}_{final} = \lambda \mathbf{v}^X + (1 - \lambda)\mathbf{v}^T. \tag{20}$$

Classification and Loss The fused representation $\mathbf{v}_{final}$ is passed through a linear layer followed by a Sigmoid activation to produce the predicted probability $\hat{p}$:

$$\hat{p} = \sigma(\mathbf{w}^\top \mathbf{v}_{final} + b), \tag{21}$$

where $\mathbf{w}$ and b are learnable parameters, and $\sigma(\cdot)$ denotes the Sigmoid function. The model is trained with standard binary cross-entropy loss:

	Chinese*(Weibo21)*			English*(Pheme)*		
	Train	Val	Test	Train	Val	Test
Real	2,784	928	928	1,098	366	366
Fake	2,692	897	897	1,183	394	394
Total	5,476	1,825	1,825	2,281	760	760

Table 1: Statistics of the datasets for Chinese and English.

$$\mathcal{L} = -\left[y \log \hat{p} + (1 - y) \log(1 - \hat{p})\right], \tag{22}$$

where $y \in \{0, 1\}$ is the ground-truth label, $y = 1$ indicates fake news and $y = 0$ indicates real news.

Experiments

Experimental Setup

In this section, we introduce the datasets and implementation details.

Datasets We evaluate PHPFND on two benchmark datasets, the Chinese-language Weibo21 (Nan et al. 2021) and the English-language Pheme (Buntain and Golbeck 2017). Weibo21 is a dataset collected from Sina Weibo, focusing on Chinese fake news. Pheme is a multi-event dataset based on tweets from Twitter, targeting English fake news detection. Both datasets are split into training, validation, and test sets with a 6:2:2 ratio. Statistics are summarized in Table 1.

Implementation Details For input-conflicting hallucination detection, thresholds ϕ_1 and τ are set to 0.62 and 0.7. For context-conflicting hallucination detection, ϕ_2, ϕ_3, and ϕ_4 are set to 0.6, 0.65 and 3. For fact-conflicting hallucination detection, K_1 and K_2 are set to 6 and 10. The maximum number of correction rounds K_3 is set to 3. Feature dimensions d and d' are set to 768 and 256. All encoders use Sentence-BERT (Reimers and Gurevych 2019) with an embedding size of 768. The LLM used in this paper is GPT-4o (Hurst et al. 2024).

Comparison and Ablation Experiments

To demonstrate the effectiveness of our method, we conduct both comparison and ablation experiments.

Baseline Methods We compare PHPFND with the following three categories of methods:

LLM-only. **1) GPT-3.5**: an LLM developed by OpenAI and supporting the popular chatbot ChatGPT (OpenAI 2022); **2) GPT-4o** (Hurst et al. 2024): a large language model with tens of billions of parameters.

SLM-only. **1) Baseline model:** a prediction model composed of a pretrained Sentence-BERT (Reimers and Gurevych 2019) and an MLP; **2) EANN** (Wang et al. 2018): a method that uses adversarial training to learn effective signals; **3) FCN-LP** (Zhao et al. 2023): a method that builds a cross-modal tweet graph and applies label propagation for multimodal fake news detection; **4) HMCAN** (Qian et al.

2021): a method that uses a hierarchical multimodal context attention network to model intra- and inter-modal relationships; **5) HSEN** (Zhang et al. 2023): a method that exploits multi-level semantic information from news images and texts.

LLM+SLM. **1) SuperICL** (Xu et al. 2023): an advanced prompt engineering approach that integrates SLMs into prompt construction; **2) ARG** (Hu et al. 2024): an adaptive reasoning guidance network that enables SLMs to selectively absorb insights from the multi-perspective reasoning of LLMs; **3) L-Defense** (Wang et al. 2024a): a method that detects fake news by simulating defense across different reasoning paths; **4) DELL** (Wan et al. 2024): a three-stage approach that integrates LLMs for news reaction generation, task explanation expansion, and expert synthesis.

Ablation Variants ***w/o ISHD:*** Removing the hallucination detection module. Statement features are directly concatenated with reasoning and passed to an MLP for prediction. ***w/o HC:*** Removing the hallucination correction while retaining hallucination detection results for downstream feature enhancement. ***w/o HDFE:*** Removing the entire feature enhancement mechanism. ***w/o AF:*** Removing the hallucination-signal-guided attention fusion and using standard attention fusion instead. ***w/o FS:*** Removing the feature separation mechanism.

Result Analysis Table 2 shows that PHPFND achieves the best overall performance across both the Weibo21 and Pheme datasets, showing strong generalization and robustness in cross-lingual fake news detection.

Baseline Comparison. Compared to *LLM-only* models, GPT-4o and GPT-3.5 fail to match PHPFND, reflecting that LLMs exhibit hallucinations in the field of fake news detection. *SLM-only* methods like HSEN perform relatively well, thanks to their multimodal fusion capacity, but lack the global reasoning and robustness provided by our hallucination modeling. Recent *LLM+SLM* hybrids improve over single-model approaches, yet still lag behind PHPFND, highlighting the effectiveness of hallucination-aware feature enhancement in PHPFND.

Ablation Study. Removing the entire hallucination detection module ($w/o\ ISHD$) causes the most significant drop, confirming its pivotal role. Removing only the correction module ($w/o\ HC$) results in a smaller but still notable decline, indicating that relying on a single detection step is insufficient.

Disabling hallucination-driven feature enhancement ($w/o\ HDFE$) causes a significant drop in macF1 score, indicating it is essential for highlighting high-risk content. Removing the risk-guided attention module ($w/o\ AF$) and feature separation strategy ($w/o\ FS$) also causes noticeable declines, confirming their contributions in distinguishing and weighting hallucination-prone regions.

Together, these results demonstrate that PHPFND's key components—including hallucination detection, correction, and signal-guided modeling—work synergistically to achieve substantial gains in robustness and accuracy across languages and detection scenarios.

Discussion on Hallucination Signals

To gain deeper insight into how hallucination signals enhance fake news detection performance, we conduct SHAP (SHapley Additive exPlanations) visualizations (Lundberg and Lee 2017) and compare the baseline model ($Base$), which directly classifies encoded features, with our model ($Ours$), which employs the feature enhancement mechanism (HDFE), for the negative class.

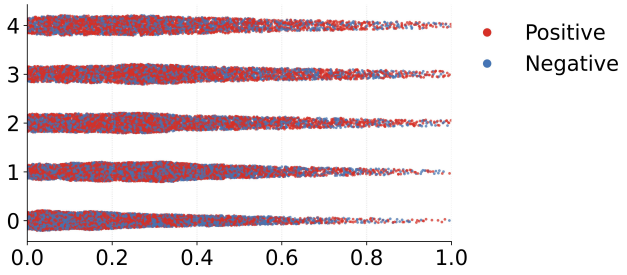

Figure 3: SHAP scatter plot on Weibo21 dataset.

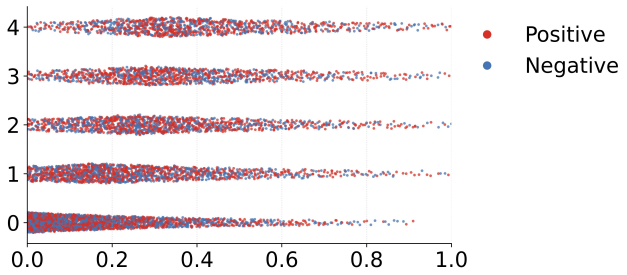

Figure 4: SHAP scatter plot on Pheme dataset.

SHAP Visualization To investigate the relationship between regions in news articles prone to inducing hallucinations in large language models and the challenging regions for fake news detectors, we group sentences by their hallucination trigger frequency (from 0 to 4) across multiple detection rounds, and apply SHAP analysis to measure each group's influence on predictions. As shown in Figures 3 and 4, where different colors indicate ground labels, the results reveal that a higher proportion of data points with larger SHAP values are found in sentences with higher hallucination frequencies. These high-risk sentences play a dominant role in shaping model output.

Misclassification Rates on Hallucinated Samples To evaluate model performance on samples that induce hallucinations in LLMs, we compare the error rates ($1 - \mathrm{ACC}$) of $Base$ and $Ours$ on the Weibo21 and Pheme datasets. As shown in Figures 5 and 6, $Ours$ significantly reduces the error rate from 32.1% to 18.7% on Weibo21 ($\downarrow$42%) and from 35.6% to 19.2% on Pheme ($\downarrow$46%). These results demonstrate that incorporating hallucination signals helps the model more effectively identify and mitigate misleading content. More importantly, they confirm that regions triggering hallucinations are not random noise, but are closely

Model	Chinese*(Weibo21)*				English*(Pheme)*			
	Acc.	**macF1**	**Pre.**	**Recall**	**Acc.**	**macF1**	**Pre.**	**Recall**
GPT-3.5	0.724	0.717	0.702	0.733	0.662	0.660	0.659	0.661
GPT-4o	0.772	0.771	0.774	0.768	0.721	0.721	0.732	0.711
Baseline	0.748	0.742	0.752	0.747	0.740	0.738	0.744	0.741
EANN	0.792	0.794	0.803	0.785	0.702	0.693	0.713	0.674
FCN-LP	0.855	0.894	0.889	0.899	0.847	0.893	0.888	0.898
HMCAN	0.874	0.878	0.861	0.896	0.864	0.835	0.832	0.838
HSEN	0.927	0.928	0.926	0.931	0.908	0.888	0.886	0.890
SuperICL	0.759	0.755	0.764	0.746	0.789	0.785	0.782	0.788
ARG	0.786	0.776	0.774	0.788	0.778	0.775	0.768	0.782
L-Defense	0.825	0.831	0.832	0.830	0.811	0.815	0.819	0.812
DELL	0.865	0.866	0.862	0.870	0.831	0.835	0.829	0.842
Ours	**0.941**	**0.942**	**0.939**	**0.944**	**0.922**	**0.925**	**0.926**	**0.924**
$w/o\ ISHD$	0.823	0.823	0.827	0.819	0.806	0.813	0.812	0.814
$w/o\ HC$	0.912	0.910	0.908	0.912	0.907	0.905	0.908	0.901
$w/o\ HDFE$	0.872	0.871	0.872	0.870	0.861	0.858	0.865	0.850
$w/o\ AF$	0.894	0.901	0.912	0.890	0.878	0.880	0.878	0.882
$w/o\ FS$	0.907	0.906	0.905	0.906	0.912	0.912	0.910	0.914

Table 2: Experimental results of performance comparison and ablation study. Bold font indicates optimal. The ablation experiment is separated from the comparison experiment by a dotted line.

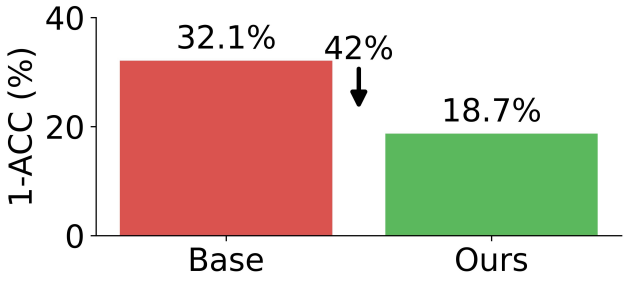

Figure 5: Misclassification rates on Weibo21 dataset.

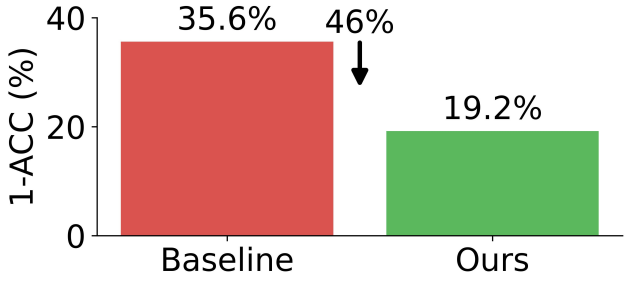

Figure 6: Misclassification rates on Pheme dataset.

tied to areas of high semantic ambiguity and decision difficulty—making them critical cues for improving the reliability and robustness of fake news detection.

Hallucination Correction Prompt

In this section, we introduce examples of prompts used in hallucination correction.

Input-Conflicting

The sentence at $< pos_1^T >$ in the LLM output $< T >$ appears unrelated to the news article $< X >$. Revise it while maintaining clear themes and instructions.

Context-Conflicting

Stance Variation. Entity $< Entity >$ shows abrupt stance changes at $< pos_2^T >$. Correct while ensuring stance consistency. *Semantic Variation.* Entity $< Entity >$ exhibits topic drift at $< pos_2^T >$. Modify while maintaining thematic coherence. *Coherence Gap.* Entity $< Entity >$ reappears at $< pos_2^T >$ after a long interval. Revise while preserving contextual connections.

Fact-Conflicting

Statements at $< pos_3^T >$ contradict real-world knowledge $< Evidence >$. Please correct with reference to verified facts while maintaining logical consistency.

Conclusion

This paper proposes a novel fake news detection framework, PHPFND, which incorporates a hallucination detector (ISHD) and a feature enhancement mechanism (HDFE). Experimental results demonstrate that PHPFND achieves superior performance on both Chinese and English benchmark datasets, significantly outperforming existing methods. Moreover, from the novel perspective that hallucinations are not merely flaws, we explore the relationship between hallucination signals and high-risk regions for fake news detectors. The results confirm that regions triggering hallucinations correspond to those significantly affecting the detector.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 6247074060), the Natural Science Foundation of Heilongjiang Province in China (No. PL2024F029), the Heilongjiang Postdoctoral Fund (LBH-Z24267) and the Basic Research Funds for Provincial Universities in Heilongjiang Province (No. 2024-KYYWF-0115).

References

Boumber, D.; Tuck, B. E.; Verma, R. M.; and Qachfar, F. Z. 2024. Llms for explainable few-shot deception detection. In *Proceedings of the 10th ACM International Workshop on Security and Privacy Analytics*, 37–47.

Buntain, C.; and Golbeck, J. 2017. Automatically identifying fake news in popular twitter threads. In *2017 IEEE international conference on smart cloud (smartCloud)*, 208–215. IEEE.

Dai, J.; Ma, J.; Zhang, W.; and Liu, Y. 2025. Graph Contrastive Adversarial Learning for Rumor Detection via Similarity-Preserving. In *ICIC*, volume 15857 of *Lecture Notes in Computer Science*, 3–14.

Dhuliawala, S.; Komeili, M.; Xu, J.; Raileanu, R.; Li, X.; Celikyilmaz, A.; and Weston, J. 2023. Chain-of-verification reduces hallucination in large language models. *arXiv preprint arXiv:2309.11495*.

Gao, L.; Madaan, A.; Zhou, S.; Alon, U.; Liu, P.; Yang, Y.; Callan, J.; and Neubig, G. 2023. Pal: Program-aided language models. In *International Conference on Machine Learning*, 10764–10799. PMLR.

Guan, T.; Liu, F.; Wu, X.; Xian, R.; Li, Z.; Liu, X.; Wang, X.; Chen, L.; Huang, F.; Yacoob, Y.; et al. 2024. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14375–14385.

Guo, H.; Zeng, W.; Tang, J.; and Zhao, X. 2023. Interpretable fake news detection with graph evidence. In *Proceedings of the 32nd ACM international conference on information and knowledge management*, 659–668.

Hu, B.; Sheng, Q.; Cao, J.; Shi, Y.; Li, Y.; Wang, D.; and Qi, P. 2024. Bad actor, good advisor: Exploring the role of large language models in fake news detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 22105–22113.

Huang, L.; Yu, W.; Ma, W.; Zhong, W.; Feng, Z.; Wang, H.; Chen, Q.; Peng, W.; Feng, X.; Qin, B.; et al. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2): 1–55.

Hurst, A.; Lerer, A.; Goucher, A. P.; Perelman, A.; Ramesh, A.; Clark, A.; Ostrow, A.; Welihinda, A.; Hayes, A.; Radford, A.; et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Ji, Z.; Lee, N.; Frieske, R.; Yu, T.; Su, D.; Xu, Y.; Ishii, E.; Bang, Y. J.; Madotto, A.; and Fung, P. 2023. Survey of hallucination in natural language generation. *ACM computing surveys*, 55(12): 1–38.

Koike, R.; Kaneko, M.; and Okazaki, N. 2024. Outfox: Llm-generated essay detection through in-context learning with adversarially generated examples. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 21258–21266.

Kumarage, T.; Bhattacharjee, A.; Padejski, D.; Roschke, K.; Gillmor, D.; Ruston, S.; Liu, H.; and Garland, J. 2023. J-guard: Journalism guided adversarially robust detection of ai-generated news. *arXiv preprint arXiv:2309.03164*.

Lin, S.; Hilton, J.; and Evans, O. 2021. Truthfulqa: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958*.

Liu, F.; Lin, K.; Li, L.; Wang, J.; Yacoob, Y.; and Wang, L. 2023. Aligning large multi-modal model with robust instruction tuning. *CoRR*.

Lundberg, S. M.; and Lee, S.-I. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.

Ma, J.; Dai, J.; Liu, Y.; Han, M.; and Ai, C. 2023. Contrastive Learning for Rumor Detection via Fitting Beta Mixture Model. In *CIKM*, 4160–4164.

Ma, J.; Liu, Y.; Han, M.; Hu, C.; and Ju, Z. 2024. Propagation Structure Fusion for Rumor Detection Based on Node-Level Contrastive Learning. *IEEE Trans. Neural Networks Learn. Syst.*, 35(12): 18649–18660.

Ma, J.; Liu, Y.; Liu, M.; and Han, M. 2022. Curriculum Contrastive Learning for Fake News Detection. In *CIKM*, 4309–4313.

Ma, J.; Zhang, L.; Liu, Y.; and Zhang, W. 2025. Multi-Task Network Guided Multimodal Fusion for Fake News Detection. In *Asian Conference on Machine Learning*, 813–828. PMLR.

Madaan, A.; Tandon, N.; Gupta, P.; Hallinan, S.; Gao, L.; Wiegreffe, S.; Alon, U.; Dziri, N.; Prabhumoye, S.; Yang, Y.; et al. 2023. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36: 46534–46594.

Nan, Q.; Cao, J.; Zhu, Y.; Wang, Y.; and Li, J. 2021. MDFEND: Multi-domain fake news detection. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, 3343–3347.

OpenAI. 2022. ChatGPT: Optimizing language models for dialogue. *https://openai.com/blog/chatgpt/*.

Qi, P.; Yan, Z.; Hsu, W.; and Lee, M. L. 2024. SNIFFER: Multimodal Large Language Model for Explainable Out-of-Context Misinformation Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13052–13062.

Qian, S.; Wang, J.; Hu, J.; Fang, Q.; and Xu, C. 2021. Hierarchical multi-modal contextual attention network for fake news detection. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, 153–162.

Radford, A.; Narasimhan, K.; Salimans, T.; Sutskever, I.; et al. 2018. Improving language understanding by generative pre-training. *https://www.mikecaptain.com/resources/pdf/GPT-1*, 1–12.

Rafailov, R.; Sharma, A.; Mitchell, E.; Manning, C. D.; Ermon, S.; and Finn, C. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36: 53728–53741.

Reimers, N.; and Gurevych, I. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Shinn, N.; Cassano, F.; Gopinath, A.; Narasimhan, K.; and Yao, S. 2023. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36: 8634–8652.

Tay, Y.; Dehghani, M.; Tran, V. Q.; Garcia, X.; Wei, J.; Wang, X.; Chung, H. W.; Shakeri, S.; Bahri, D.; Schuster, T.; et al. 2022. Ul2: Unifying language learning paradigms. *arXiv preprint arXiv:2205.05131*.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

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

Wang, B.; Ma, J.; Lin, H.; Yang, Z.; Yang, R.; Tian, Y.; and Chang, Y. 2024a. Explainable fake news detection with large language model via defense among competing wisdom. In *Proceedings of the ACM Web Conference 2024*, 2452–2463.

Wang, B.; Wu, F.; Han, X.; Peng, J.; Zhong, H.; Zhang, P.; Dong, X.; Li, W.; Li, W.; Wang, J.; et al. 2024b. Vigc: Visual instruction generation and correction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 5309–5317.

Wang, W.-Y.; Chang, Y.-C.; and Peng, W.-C. 2024. Stylenews: Incorporating stylized news generation and adversarial verification for neural fake news detection. *arXiv preprint arXiv:2401.15509*.

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

Xu, C.; Xu, Y.; Wang, S.; Liu, Y.; Zhu, C.; and McAuley, J. 2023. Small models are valuable plug-ins for large language models. *arXiv preprint arXiv:2305.08848*.

Yang, R.; Gao, W.; Ma, J.; Lin, H.; and Wang, B. 2024. Reinforcement tuning for detecting stances and debunking rumors jointly with large language models. *arXiv preprint arXiv:2406.02143*.

Yin, S.; Fu, C.; Zhao, S.; Xu, T.; Wang, H.; Sui, D.; Shen, Y.; Li, K.; Sun, X.; and Chen, E. 2024. Woodpecker: Hallucination correction for multimodal large language models. *Science China Information Sciences*, 67(12): 220105.

Zeng, F.; and Gao, W. 2023. Prompt to be consistent is better than self-consistent? few-shot and zero-shot fact verification with pre-trained language models. *arXiv preprint arXiv:2306.02569*.

Zhang, Q.; Liu, J.; Zhang, F.; Xie, J.; and Zha, Z.-J. 2023. Hierarchical Semantic Enhancement Network for Multimodal Fake News Detection. In *Proceedings of the 31st ACM International Conference on Multimedia*, 3424–3433.

Zhang, X.; and Gao, W. 2023. Towards llm-based fact verification on news claims with a hierarchical step-by-step prompting method. *arXiv preprint arXiv:2310.00305*.

Zhang, Y.; Li, Y.; Cui, L.; Cai, D.; Liu, L.; Fu, T.; Huang, X.; Zhao, E.; Zhang, Y.; Chen, Y.; et al. 2025. Siren's song in the ai ocean: A survey on hallucination in large language models. *Computational Linguistics*, 1–45.

Zhao, W.; Nakashima, Y.; Chen, H.; and Babaguchi, N. 2023. Enhancing Fake News Detection in Social Media via Label Propagation on Cross-modal Tweet Graph. In *Proceedings of the 31st ACM International Conference on Multimedia*, 2400–2408.