FACTGUARD: Event-Centric and Commonsense-Guided Fake News Detection

Jing He[1], Han Zhang[1], Yuanhui Xiao[1], Wei Guo[1], Shaowen Yao[1], Renyang Liu[2,*]

[1]School of Software and AI, Yunnan University
[2]Institute of Data Science, National University of Singapore
hejing@ynu.edu.cn,{hanzhang, xiaoyuanhui, guoweiynu}@stu.ynu.edu.cn, yaosw@ynu.edu.cn, ryliu@nus.edu.sg

Abstract

Fake news detection methods based on writing style have achieved remarkable progress. However, as adversaries increasingly imitate the style of authentic news, the effectiveness of such approaches is gradually diminishing. Recent research has explored incorporating large language models (LLMs) to enhance fake news detection. Yet, despite their transformative potential, LLMs remain an untapped goldmine for fake news detection, with their real-world adoption hampered by shallow functionality exploration, ambiguous usability, and prohibitive inference costs. In this paper, we propose a novel fake news detection framework, dubbed FACTGUARD, that leverages LLMs to extract event-centric content, thereby reducing the impact of writing style on detection performance. Furthermore, our approach introduces a dynamic usability mechanism that identifies contradictions and ambiguous cases in factual reasoning, adaptively incorporating LLM advice to improve decision reliability. To ensure efficiency and practical deployment, we employ knowledge distillation to derive FACTGUARD-D, enabling the framework to operate effectively in cold-start and resource-constrained scenarios. Comprehensive experiments on two benchmark datasets demonstrate that our approach consistently outperforms existing methods in both robustness and accuracy, effectively addressing the challenges of style sensitivity and LLM usability in fake news detection.

Code — https://github.com/ryliu68/FACTGUARD
Extended version — https://arxiv.org/abs/2511.10281

1 Introduction

Social media platforms have become dominant channels for information dissemination, surpassing traditional media in both scale and societal influence. However, the lack of effective content moderation on these platforms has facilitated the rapid spread of sensational fake news, further amplified by recommendation algorithms. Consequently, platforms such as Sina-Weibo and Facebook have faced significant challenges related to misinformation propagation (Refutation 2021; Avaaz 2020). Prior studies have demonstrated that the widespread circulation of fake news during major public events can trigger social panic and disrupt governance (Grinberg et al. 2019; Zhang et al. 2024; Bursztyn et al. 2020). Given the overwhelming volume of online information, manual verification of news authenticity is infeasible, which underscores the necessity of developing automated fake news detection methods based on advanced techniques, such as deep learning, with a particular emphasis on early identification to curb the spread and societal impact of misinformation.

Early efforts in this direction have predominantly relied on insights from psychology and linguistics. In particular, a substantial body of research is rooted in psychological theories, such as the Undeutsch hypothesis (Amado, Arce, and Fariña 2015), emphasizing linguistic style differences between truthful and deceptive statements. Stylometric and emotion-based methods (Potthast et al. 2018; Rashkin et al. 2017; Ajao, Bhowmik, and Zargari 2019; Giachanou, Rosso, and Crestani 2019; Zhu et al. 2022) have thus become mainstream approaches for identifying fake news. However, as these methods primarily capture superficial features, recent studies have shown that adversaries can effectively evade detection by imitating the writing style of authentic news(Wu, Guo, and Hooi 2024). This reliance on surface-level cues makes existing systems highly vulnerable to news text writing style.

To address these limitations, the research community has increasingly explored leveraging large language models (LLMs) for fake news detection. For example, some studies employ LLMs to generate adversarial samples with diverse writing styles to enhance model robustness (Wu, Guo, and Hooi 2024; Wang et al. 2024). Other approaches integrate multiple perspectives, such as combining style detection with commonsense reasoning (Hu et al. 2024), constructing multi-agent debate frameworks to aggregate different viewpoints (Liu et al. 2025) or using LLMs to simulate different news readers to generate diverse comments (Nan et al. 2024). Despite these methodological advances, several critical challenges remain unresolved. In particular, the effectiveness of style-based sample generation methods against previously unseen style attacks remains uncertain. LLM-based detection models, though promising, often suffer from low accuracy in few-shot and chain-of-thought reasoning scenarios (Hu et al. 2024), are prone to hallucinations (Xu, Jain, and Kankanhalli 2024). Besides, some methods fo-

*Corresponding author
Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

cus only on correctly judged samples during training, while the correctness of such judgments remains unknown during inference (Hu et al. 2024), resulting in the lack of reliable mechanisms for usability assessment. Moreover, multi-agent debate frameworks and role-playing–based comment generation typically incur considerable computational and time costs, which limits their practicality in cold-start[1] and resource-constrained environments[2] (Liu et al. 2025).

To bridge this gap, we propose the News Extracted Topic-Content and Commonsense Rationale Model (FACTGUARD), a comprehensive framework for robust fake news detection. FACTGUARD leverages the semantic understanding capabilities of LLMs to extract event-centric information, thereby reducing the influence of textual style. By integrating LLM-generated commonsense reasoning with advanced content extraction and dynamic reliability assessment, FACTGUARD enables a more accurate assessment of news veracity. Furthermore, the framework incorporates knowledge distillation, thereby supporting practical deployment across resource-rich[3], cold-start, and resource-constrained settings and achieving a balance between accuracy and efficiency.

Specifically, in journalism communication theory, news come from the real-world events (Galtung and Ruge 1965) and style rewriting is a common deceptive strategy in fake news (Potthast et al. 2018) to accelerate news dissemination. So FACTGUARD first utilizes LLMs with carefully designed prompts to extract the core topic and principal content from news articles. This process filters out stylistic noise and preserves essential event information. To ensure the quality and relevance of the extracted content, we introduce a two-stage constraint mechanism: a text similarity metric is applied during extraction to maintain consistency with the original news, followed by an information density metric that evaluates informativeness post-extraction. The resulting event-centric content, being more objective and concise, is then semantically compared with LLM-generated commonsense rationale to enhance fake news detection. To further improve detection reliability, FACTGUARD incorporates an LLM Rationale usability module that treats the LLM as an advisor and dynamically assesses the trustworthiness of its advice via a dual-branch structure. One branch adaptively controls the influence of LLM-based judgments, while the other emphasizes potential conflicts or ambiguities identified through commonsense reasoning, enabling the model to better address complex cases. In addition, to support deployment in cost- and efficiency-sensitive scenarios, we introduce a knowledge distillation scheme (Hinton, Vinyals, and Dean 2015). This mechanism transfers knowledge from the full FACTGUARD model to a lightweight variant, FACTGUARD-D, which delivers efficient inference while preserving strong detection performance.

Extensive experiments on two widely used real-world fake news detection datasets, GossipCop (Shu et al. 2020) and Weibo21 (Nan et al. 2021), demonstrate that FACTGUARD consistently outperforms state-of-the-art baselines across multiple key metrics, including accuracy and robustness. The distilled variant, FACTGUARD-D, also achieves competitive results with minimal performance degradation, confirming its practicality and robustness in resource-constrained settings. Our contributions are as follows:

- We propose FACTGUARD, a novel framework that effectively mitigates the impact of textual style on fake news detection and enables robust integration of LLM-based reasoning. To accommodate diverse practical needs, FACTGUARD is suitable for resource-rich scenarios, whereas its distilled variant, FACTGUARD-D, is tailored for cold-start and resource-constrained scenarios.
- We develop an LLM-based news extraction approach that leverages semantic understanding to obtain key topics and event content, thereby reducing style interference and enabling alignment with commonsense reasoning.
- We introduce an LLM rationale usability module, which dynamically adjusts the influence of LLM advice through a dual-branch structure based on their reliability and the presence of commonsense conflicts, ensuring the effective and adaptive use of LLM knowledge.
- We conduct extensive experiments on the GossipCop and Weibo21 datasets, which demonstrate the effectiveness and efficiency of the proposed FACTGUARD and FACTGUARD-D models.

2 Related Work

2.1 Traditional Fake News Detection

Fake news detection focuses on the early identification of misinformation, primarily based on the textual content available at publication (Qian et al. 2018). Early methods mainly relied on machine learning models with handcrafted features, including keywords, grammatical errors (Granik and Mesyura 2017), shallow linguistic patterns (Wang 2017), and statistical cues such as text length, capitalization, and punctuation (Castillo, Mendoza, and Poblete 2011). With advances in deep learning, LSTM-based approaches were introduced to capture linguistic differences in news with satirical or rumor styles (Rashkin et al. 2017), and some studies explored sentiment-related features (Ajao, Bhowmik, and Zargari 2019; Giachanou, Rosso, and Crestani 2019). However, these methods fundamentally rely on surface-level features, making them vulnerable to variations in writing style.

To address these limitations, style modeling with pre-trained language models such as BERT and RoBERTa (Przybyla 2020) has become common in fake news detection. Nevertheless, approaches relying solely on textual features remain inadequate for combating increasingly sophisticated misinformation. Recent studies have therefore incorporated background knowledge beyond textual content, including

[1]Cold-start refers to the early-stage fake news detection setting where only the news content is available. In this setting, high inference efficiency is required.

[2]Resource-constrained refers to settings where LLMs cannot be accessed or invoked, and the model can only rely on news content for inference in such cases.

[3]Resource-rich refers to settings where LLMs are available and can be employed without restrictions, enabling improved fake news detection performance.

social context (Shu et al. 2019; Cui et al. 2022), social emotion (Zhang et al. 2021), news environment (Sheng et al. 2022), and external knowledge (Hu et al. 2022). While small language models (SLMs) offer certain improvements, their limited knowledge and capacity continue to constrain further progress in fake news detection.

2.2 LLM-based Fake News Detection

Recent studies have leveraged the strong language generation and comprehension capabilities of LLMs for content generation and enhancement in fake news detection. For content generation, LLM-Fake analyzes the psychological motivations behind LLM-generated fake news and constructs the MegaFake dataset (Wang et al. 2024); SheepDog employs LLM-generated multi-style samples as adversarial data to improve detector robustness (Wu, Guo, and Hooi 2024). For content enhancement, ARG leverages prompt engineering to guide LLM in multi-perspective analysis, with SLM integrating the final judgment (Hu et al. 2024). LEKD incorporates offline LLM knowledge via semantic graph alignment and knowledge distillation (Chen et al. 2025). TED introduces a structured multi-agent debate mechanism, enabling LLM to reason from diverse perspectives (Liu et al. 2025). GenFEND simulate readers with different identities to generate diverse comments, thereby providing additional information for early-stage detection (Nan et al. 2024).

Despite these advances, the fundamental problem of style sensitivity remains unresolved. While LLM-based content generation increases sample diversity, it often fails to capture event semantics and does not fully eliminate style-related interference (Wang et al. 2024; Wu, Guo, and Hooi 2024). In addition, augmented data generated by LLMs is often not effectively integrated with the detection backbone, resulting in limited overall improvement (Hu et al. 2024). Excessive reliance on LLMs and their high inference costs further constrain practical deployment (Hu et al. 2024; Liu et al. 2025; Nan et al. 2024), particularly in cold-start or resource-constrained scenarios.

In summary, vulnerability to writing style, insufficient event-centric modeling, and the challenge of efficiently integrating LLM capabilities remain open problems. Motivated by these gaps, this paper proposes a new framework, FACTGUARD, that systematically addresses style interference, leverages LLMs' commonsense reasoning capacity, and supports efficient deployment for fake news detection.

3 Framework

3.1 Preliminary

Problem Formulation. In resource-rich scenarios, we consider a dataset $\mathcal{D}_{\text{news}} = \{(n_i, c_i, r_i)\}_{i=1}^{N}$, where n_i denotes the news text, while c_i and r_i represent the event-based topic-content and the commonsense rationale extracted by an LLM. The goal for each news item $D_i = (n_i, c_i, r_i)$ is to predict $\hat{y}_i \in \{0, 1\}$, where 1 indicates fake news and 0 indicates true news, by integrating the original text with LLM-generated features. The detection model (i.e., FACTGUARD) encodes n_i via a dual-attention module $f_\theta(n_i)$, fuses c_i and r_i with cross-attention and usability assessment $g_\phi(\mathrm{CA}(c_i, r_i))$, and concatenates both for final prediction:

$$\hat{y}_i = \mathrm{MLP}\left([f_\theta(n_i);\, g_\phi(\mathrm{CA}(c_i, r_i))]\right), \tag{1}$$

where [;] denotes vector concatenation. To further address cold-start and resource-constrained scenarios, we distill an efficiency and resource-friendly student model (i.e., FACTGUARD-D) from the teacher model FACTGUARD, where the FACTGUARD-D only takes n_i as input and predicts its label $\hat{y}_i$:

$$\hat{y}_i = f_\psi(n_i). \tag{2}$$

3.2 FACTGUARD Overview

Figure 1(a–c) illustrates the main modules as well as the training and inference procedures of FACTGUARD. The core goal of FACTGUARD is to achieve style debiasing and fully exploit the capabilities of LLMs for robust fake news detection. For each news item n, the model first employs the LLM to extract topic-content information c as well as commonsense rationale r. These elements, together with the original news text, are encoded using an SLM (see Figure 1(a)). The Topic-Content&Rationale Interactor enables deep feature interaction between the extracted topic-content and the commonsense rationale, while the Rationale Usability Evaluator adaptively assigns weights to the LLM-provided advice. The resultant interacted features f_{llm} are then aggregated with the news features f_N. The fused representations are utilized for veracity prediction for n (see Figure 1(b)). During training, three loss functions—L_{cls}, $L_{usability}$, and L_{text}—are employed to optimize model parameters. Once FACTGUARD is well trained, it can be used to predict the veracity of the unseen news sample n by leveraging the inputs c, r, and n (see Figure 1(c)).

3.3 Feature Extraction

Pretrained SLMs such as BERT or RoBERTa are trained on large-scale datasets in an unsupervised fashion, enabling them to generate high-dimensional contextual representations well-suited for various downstream tasks. To effectively extract information features, we employ these models as text encoders within our framework. Specifically, for a given news item n, the extracted topic-content c, and the commonsense rationale r, we denoted them as N (news), C (topic-content), and R (commonsense rationale), respectively.

3.4 Feature Concatenation

This module aims to obtain high-quality representations for both the LLM-generated augmented information and the original news content, facilitating their effective integration and collaboration as the foundation for fake news detection. Feature integration is performed via concatenation of the respective representations, enabling subsequent modules to leverage comprehensive contextual information.

Topic-Content&Rationale Interactor. To enable comprehensive feature exchange between the LLM-extracted topic-content and commonsense rationale, we introduce a

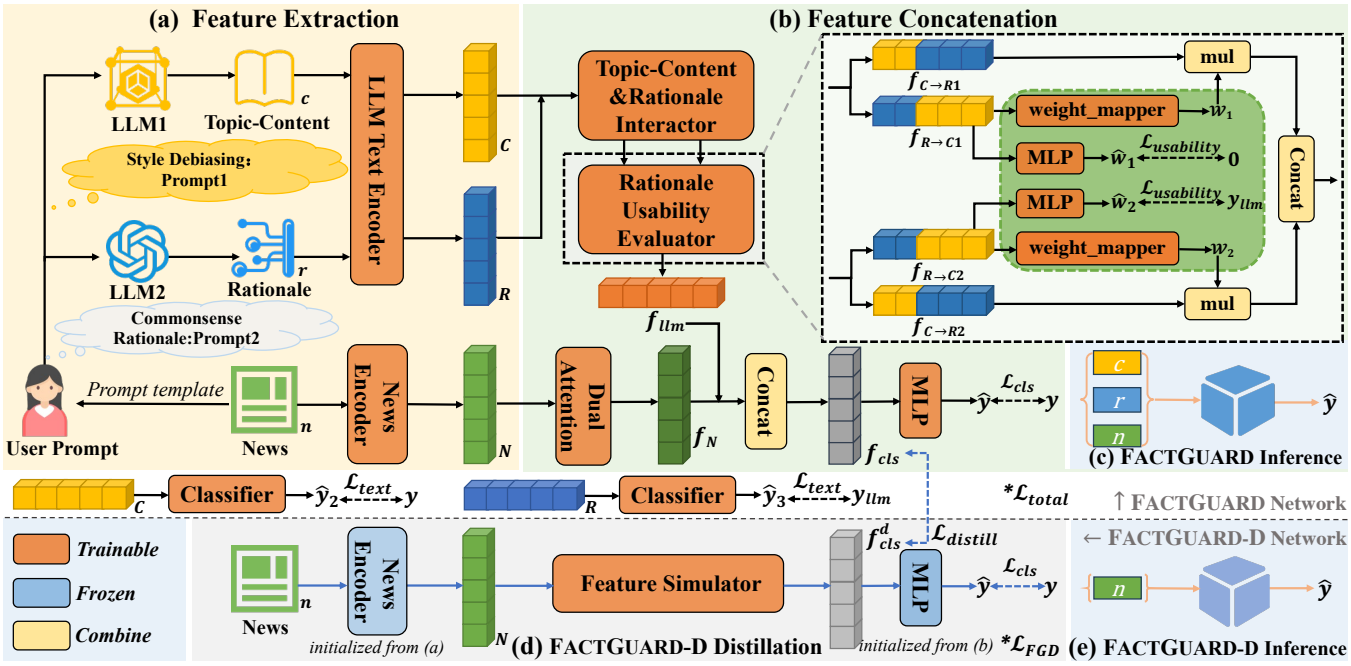

Figure 1: Overview of FACTGUARD and FACTGUARD-D. FACTGUARD main consists of two modules: **(1) Feature Extraction**, which identifies topic content and enables commonsense reasoning for each news article using an LLM. The resulting features and the original text are encoded for downstream processing. **(2) Feature Concatenation**, which adaptively integrates LLM-derived features with news content via a cross-attention mechanism and the Rationale Usability Evaluator, followed by MLP-based classification. After training, knowledge distillation yields a lightweight FACTGUARD-D without LLMs' advice.

dual cross-attention module based on multi-head attention. The computation is formulated as follows:

$$\text{head}_i = \text{softmax}\left(\frac{Q_i K_i^\top}{\sqrt{d_k}}\right) V_i, \tag{3}$$

$$\text{CA}(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h) W^O, \tag{4}$$

where $Q_i = QW_i^Q$, $K_i = KW_i^K$, and $V_i = VW_i^V$. Here, d_k is the dimension of each attention head, h is the number of heads, and W^O is the output projection matrix. Given topic-content c and commonsense rationale r, after embedding, the interactions are computed as:

$$f_{C\to R} = \text{AvgPool}\left(\text{CA}(C, R, R)\right), \tag{5}$$

$$f_{R\to C} = \text{AvgPool}\left(\text{CA}(R, C, C)\right), \tag{6}$$

where $\text{AvgPool}(\cdot)$ denotes average pooling applied over the token representations output by the cross-attention layer. $f_{C\to R}$ represents the LLM advice feature vector, and $f_{R\to C}$ serves as a weighting factor in the Rationale Usability Evaluator.

Rationale Usability Evaluator. Directly enforcing consistency between LLM judgments and ground-truth labels may result in the loss of valuable complementary features provided by the LLMs. To address this, we propose a rationale usability evaluation module to dynamically adjust the fusion weights of LLM features, thereby maximizing the utility of LLM-generated knowledge. This module adopts a dual-branch MLP structure: one branch reduces its contribution as the LLMs' direct detection capability is limited, while the other increases its contribution as commonsense reasoning can provide more effective information when it identifies contradictions or uncertainty. A three-layer MLP (weight_mapper) maps the feature vectors $f_{R\to C_i}$ to fusion weights w_i as follows:

$$w_i = \text{sigmoid}(\text{weight\_mapper}(f_{R\to C_i})), \quad i = 1, 2. \tag{7}$$

The final LLM feature representation f_{llm} is then computed as:

$$f_{llm} = [w_1 \cdot f_{C\to R1}; w_2 \cdot f_{C\to R2}], \tag{8}$$

where w_1 and w_2 are the fusion weights for the two branches, and $f_{C\to R1}$, $f_{C\to R2}$ denote their respective interaction features.

Dual Attention Fusion. To further improve the expressiveness and robustness of features derived from BERT or RoBERTa, we introduce a linear attention mechanism that adaptively assigns higher weights to salient tokens, thereby suppressing irrelevant or noisy information:

$$\text{Attn}(X) = \sum_{t=1}^{T} \text{softmax}(Wx_t + b) \cdot x_t, \tag{9}$$

where x_t denotes the input feature at position t, W and b are learnable parameters, and T is the token sequence length. To enhance model robustness, a dual-branch architecture is adopted, where the same linear attention module is applied in parallel to both branches. The outputs of the two branches

are then averaged to obtain the final news feature representation:

$$f_N = \frac{\text{Attn}(N) + \text{Attn}(N)}{2}. \tag{10}$$

Feature Concatenation. Based on the outputs obtained in the previous step, the news feature vector f_N and the LLM-enhanced feature vector f_{llm} are summed to facilitate the final prediction. For each news item n with label $y \in \{0, 1\}$, these vectors are combined to produce the final feature representation f_{cls}, computed as:

$$f_{cls} = [f_N; f_{llm}]. \tag{11}$$

f_{cls} is subsequently input into an MLP classifier to predict the veracity label:

$$\hat{y} = \text{MLP}(f_{cls}). \tag{12}$$

3.5 Training

Data Process. To enhance semantic understanding and reduce the influence of writing style, LLMs is leveraged to extract the topic-content of each news article. Additionally, commonsense rationale analysis is performed by the LLMs to identify and judge content that may contradict commonsense.

Objective. The FACTGUARD method is designed with three principal objectives: ① to achieve accurate prediction of news veracity; ② to effectively integrate model recommendations and fully leverage the capabilities of LLMs; and ③ to enhance the representation of information augmentation provided by LLMs. Accordingly, the overall objective loss function is defined as a weighted sum of the prediction loss, the LLM rationale usability loss, and the information augmentation representation loss.

To improve the final detection performance, the Binary Cross-Entropy (BCE) classification loss is computed to guide the model in accurately identifying fake news:

$$\mathcal{L}_{cls} = \text{BCE}(\hat{y}, y). \tag{13}$$

To supervise the learning of LLM features, the supervision signals for the weights are set as 0 for one branch and y_{llm} for the other:

$$\hat{w}_i = \text{sigmoid}(\text{MLP}(f_{C \to R_i})), \quad i = 1, 2, \tag{14}$$

$$\mathcal{L}_{usability} = \text{BCE}(\hat{w}_1, 0) + \text{BCE}(\hat{w}_2, y_{llm}). \tag{15}$$

To enhance the utility of LLM-generated augmentations, we employ an auxiliary task that aligns extracted semantics and commonsense reasoning with ground-truth labels and LLM veracity judgments. Augmented representations are fed into a classifier composed of a linear attention and an MLP head (without sigmoid), optimized by cross-entropy (CE) loss:

$$\hat{y}_2 = \text{Classifier}(C), \quad \hat{y}_3 = \text{Classifier}(R), \tag{16}$$

$$\mathcal{L}_{text} = \text{CE}(\hat{y}_2, y) + \text{CE}(\hat{y}_3, y_{llm}). \tag{17}$$

The total loss function is a weighted sum of the aforementioned terms:

$$\mathcal{L}_{total} = \mathcal{L}_{cls} + \alpha \frac{\mathcal{L}_{usability}}{2} + \beta \frac{\mathcal{L}_{text}}{2}, \tag{18}$$

where α and β are hyperparameter weights, and the division by 2 is used to average the two sub-losses, ensuring balanced contributions from each component in the overall loss.

3.6 Inference

In resource-rich scenarios, LLMs are leveraged via prompt engineering to extract the topic-content, and commonsense rationales of news articles. The extracted outputs, together with the original news text, are subsequently fed into the well-trained frozen FACTGUARD model for veracity prediction.

3.7 FACTGUARD-D

Distillation. Directly invoking LLMs for each prediction in FACTGUARD is impractical in resource-constrained or latency-sensitive cold-start scenarios due to the substantial overhead of real-time LLM prompting for text extraction and commonsense reasoning. To address this, we develop a llm-free student model via knowledge distillation from the trained FACTGUARD, following a teacher–student paradigm (Hu et al. 2024). The core idea is to transfer and internalize the teacher model's reasoning knowledge into a parameterized lightweight student network. Specifically, as illustrated in Figure 1(d), the student model's news encoder and classifier are initialized from the trained FACTGUARD. To acquire the teacher's reasoning capabilities, a feature simulator is implemented as a four-layer Transformer encoder and a linear attention module to internalize the teacher's knowledge. In addition to the standard prediction loss $\mathcal{L}_{cls}$ as in FACTGUARD, the student model is further supervised by a feature distillation loss $\mathcal{L}_{distill}$, which encourages the student's feature representation f^d_{cls} to approximate that of the teacher f_{cls} by minimizing the mean squared error (MSE) between them:

$$\mathcal{L}_{distill} = \text{MSE}(f^d_{cls}, f_{cls}). \tag{19}$$

Inference. In cold-start and resource-constrained scenarios, the FACTGUARD-D model operates exclusively on the original news text n, achieving fast predictions with only a slight reduction in accuracy.

4 Experiments

This section presents comprehensive experimental studies of the proposed FACTGUARD and FACTGUARD-D models. We first introduce the experimental setup . Subsequently, we compare FACTGUARD with a wide range of baselines, conduct ablation studies to assess the contribution of each model component, analyze parameter sensitivity, and discuss challenges associated with LLM-based text extraction.

4.1 Setup

Datasets. We employ the Weibo21 (Chinese) (Nan et al. 2021) and GossipCop (English) (Shu et al. 2020) for evaluation. Both datasets are preprocessed by deduplication and temporal splitting, following established practices (Zhu et al. 2022; Mu, Bontcheva, and Aletras 2023; Hu et al. 2024), to mitigate the risk of data leakage and prevent overestimation of SLM performance. In addition, we also utilize the commonsense rationales from (Hu et al. 2024).

Group	Model	Weibo21				GossipCop			
		macF1	Acc.	$F1_{real}$	$F1_{fake}$	macF1	Acc.	$F1_{real}$	$F1_{fake}$
G1	GPT-3.5-turbo*	0.725	0.734	0.774	0.676	0.702	0.813	0.884	0.519
	GPT-4o-mini#	0.725	0.746	0.780	0.670	0.691	0.845	0.909	0.472
	ChatEval-o#	0.694	0.717	0.778	0.611	0.733	0.860	0.919	0.546
	ChatEval-s#	0.694	0.719	0.780	0.608	0.738	0.869	0.923	0.553
G2	BERT*	0.753	0.754	0.769	0.737	0.765	0.862	0.916	0.615
	RoBERTa	0.753	0.755	0.775	0.731	0.765	0.862	0.916	0.613
	EANN*	0.754	0.756	0.773	0.736	0.763	0.864	0.918	0.608
	Publisher-Emo*	0.761	0.763	0.784	0.738	0.766	0.868	0.920	0.611
	ENDEF*	0.765	0.766	0.779	0.751	0.768	0.865	0.918	0.618
G3	Bert + Rationale*	0.767	0.769	0.787	0.748	0.777	0.870	0.921	0.633
	SuperICL*	0.757	0.759	0.779	0.734	0.736	0.864	0.920	0.551
	Bert + GenFEND	0.755	0.760	0.791	0.719	0.764	0.875	0.926	0.603
	Roberta + GenFEND	0.771	0.774	0.796	0.747	0.770	0.866	0.919	0.621
	ARG*	0.784	0.786	0.804	0.764	0.790	0.878	0.926	0.653
	TED#	<u>0.795</u>	<u>0.798</u>	<u>0.815</u>	<u>0.774</u>	<u>0.803</u>	**0.892**	0.932	<u>0.674</u>
	Ours	**0.801**	**0.804**	**0.824**	**0.777**	**0.805**	**0.892**	**0.935**	**0.675**
G4	ARG-D*	0.771	0.772	0.785	0.756	0.778	0.870	0.921	0.634
	Ours	0.788	0.790	0.807	0.769	0.790	0.888	<u>0.933</u>	0.647

Table 1: Performance comparison on Weibo21 and GossipCop datasets across four metrics, i.e.,macF1, Accuracy, $F1_{real}$, and $F1_{fake}$. The highest result in each category is **bolded** and the second highest result is <u>underlined</u>. In the results table, * means the result is from (Hu et al. 2024) and # means the result is from (Liu et al. 2025).

Baselines. Recent fake news detection methods predominantly rely on LLMs and SLMs, and can be categorized into four groups. Among them, we involved 14 representative baselines in this work. The first group (G1) comprises LLM-only methods, including GPT-3.5-turbo (OpenAI 2023), GPT-4o-mini (OpenAI 2024), ChatEval-o (one-by-one strategy) (Chan et al. 2024), and ChatEval-s (Simultaneous-Talk strategy) (Chan et al. 2024). The second group (G2) consists of SLM-only methods, such as BERT (Devlin et al. 2019), RoBERTa (Liu et al. 2019), EANN (Wang et al. 2018), Publisher-Emo (Zhang et al. 2021), and ENDEF (Zhu et al. 2022). The third group (G3) includes LLM-SLM methods, such as BERT + Rationale (Hu et al. 2024), SuperICL (Zhong et al. 2023), ARG (Hu et al. 2024), BERT + GenFEND(Nan et al. 2024), RoBERTa + GenFEND and TED (Liu et al. 2025). The fourth group (G4) comprises methods employing model distillation, such as ARG-D (Hu et al. 2024).

Metrics. We evaluate performance using four metrics: Accuracy ($\mathrm{Acc.}$), $\mathrm{F1_{real}}$, $\mathrm{F1_{fake}}$, and Macro-F1 ($\mathrm{macF1}$).

Implementation Details. We utilize bert-base-chinese[4] (Devlin et al. 2019) as the text encoder for Chinese FACTGUARD model and roberta-base[5] (Liu et al. 2019) for English FACTGUARD model. For the Weibo21 and GossipCop dataset, topic-content extraction is performed using locally deployed DeepSeek-R1-Distill-Llama-8B[6] (DeepSeek-AI 2025) and SOLAR-10.7B-Instruct-v1.0-uncensored[7] (Kim et al. 2024) (Kim et al. 2024), respectively. Commonsense reasoning modules for both datasets are adopted from prior work (Hu et al. 2024). We employ the AdamW optimizer with an initial learning rate of $2e-4$ and a weight decay of $5e-5$. Early stopping with a patience of 5 epochs is applied to prevent overfitting. All experiments are conducted on a single NVIDIA A100 (40GB) GPU with a fixed random seed of 3759, PyTorch version 1.13.0.

4.2 Comparative Results

To evaluate the effectiveness of the proposed FACTGUARD model, we conduct systematic experiments on the Weibo21 and GossipCop datasets. As shown in Table 1, FACTGUARD consistently outperforms all baseline methods across multiple evaluation metrics, achieving the best results across both datasets. These results underscore the superior cross-lingual generalization and stability of the proposed model.

Weibo21. On the Weibo21 dataset, FACTGUARD outperforms the strongest baseline TED by 0.8% in accuracy and 0.9% in $F1_{real}$, along with notable macF1 gains. These improvements arise from its LLM-driven topic–content extraction, dual-branch reasoning, and efficient rationale usability module, which jointly combine SLM efficiency with LLM reasoning capacity. Compared with TED's complex multi-agent debate framework, FACTGUARD achieves higher accuracy with only two simple prompts, significantly reduc-

[4]https://huggingface.co/google-bert/bert-base-chinese

[5]https://huggingface.co/FacebookAI/roberta-base

[6]https://huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Llama-8B

[7]https://huggingface.co/upstage/SOLAR-10.7B-Instruct-v1.0

Group	Model	Weibo21				GossipCop			
		macF1	Acc.	$F1_{real}$	$F1_{fake}$	macF1	Acc.	$F1_{real}$	$F1_{fake}$
G1	**FACTGUARD**	**0.801**	**0.804**	**0.824**	**0.777**	**0.805**	**0.892**	**0.935**	**0.675**
G2	News	0.768	0.769	0.784	0.751	0.765	0.862	0.916	0.613
	Topic-Content	0.690	0.691	0.708	0.672	0.769	0.861	0.915	0.624
	Commonsense	0.678	0.685	0.728	0.627	0.698	0.832	0.899	0.498
G3	w/o News	0.718	0.722	0.753	0.683	0.773	0.873	0.924	0.623
	w/o Topic-Content	0.772	0.774	0.794	0.750	0.779	0.875	0.925	0.632
	w/o Commonsense	0.770	0.773	0.797	0.743	0.779	0.871	0.922	0.633
G4	w/o llm-usability	0.778	0.780	0.794	0.763	0.780	0.878	0.927	0.633
	use ARG-usefulness	0.782	0.782	0.793	0.770	0.775	0.872	0.923	0.628

Table 2: An ablation study of FACTGUARD on the Weibo21 and GossipCop datasets evaluates the contribution of each model component. Specifically, G2 assesses the predictive power of individual input features (original news, LLM-extracted topic-content, LLM commonsense rationale judgment); G3 quantifies the impact of ablating each core input module; and G4 investigates the effects of removing LLM rationale usability module on overall performance and cross-lingual generalization.

ing computational resource and inference costs. The distilled variant, FACTGUARD-D, further surpasses ARG and ARG-D, demonstrating effective compression and strong practical performance.

GossipCop. On the GossipCop dataset, although the improvements are smaller, FACTGUARD still attains the best macF1 (0.805) and $F1_{real}$ (0.935). Under limited computational resources, FACTGUARD-D also achieves lower misclassification rates, confirming its broad applicability. The performance differences across datasets mainly result from class imbalance (more fake news in Weibo21, more real news in GossipCop), yet FACTGUARD remains stable and generalizable across languages.

In summary, both FACTGUARD and FACTGUARD-D demonstrate strong performance, cross-lingual robustness, and model compression capability, making them efficient and reliable solutions for multilingual fake news detection.

4.3 Ablation Study

We conduct ablation studies on the GossipCop and Weibo21 datasets to evaluate the effectiveness of each module in the proposed FACTGUARD framework, focusing on the LLM-based topic-content extraction, commonsense rationale judgment, and usability evaluation modules. The ablation experiments are organized into three groups:

As shown in Table 2, the main findings are as follows: (1) Removing the original news representation yields the largest reduction in overall performance, confirming its indispensable role as the foundation for fake news detection. (2) Removing the LLM-based topic-content extraction module results in a notable drop in macF1 and $F1_{real}$, highlighting its importance in capturing essential event information and reducing stylistic noise. (3) Excluding the LLM commonsense rationale module further degrades performance, demonstrating its value in improving factual consistency and reasoning. (4) Omitting the usability evaluation module or use ARG's LLM usefulness module also leads to decreased performance, underscoring its role in aligning LLM inference with robust news representations. (5) The LLM-extracted topic-content module and the LLM commonsense rationale module need to be used in conjunction to achieve the maximum performance improvement. Overall, these results validate that each component is essential for FACTGUARD 's strong performance in multilingual fake news detection.

In addition to the main ablation experiments, we performed both grid search and random search to determine optimal loss weight parameters for the multi-objective loss functions in the FACTGUARD model. For the Weibo21 configuration, the best results were achieved with $\alpha = 0.40$ and $\beta = 0.16$, while for the GossipCop configuration, the optimal values were $\alpha = 0.50$ and $\beta = 0.58$. To facilitate effective distillation, we set the distillation coefficient λ in the loss function to 8 for both the Chinese and English models in FACTGUARD-D.

5 Conclusion

We propose FACTGUARD, a model that leverages LLMs for semantic understanding and commonsense reasoning to improve fake news detection performance. By extracting topic and core content and employing a usability evaluation module in commonsense rationale, FACTGUARD effectively reduces style bias and integrates LLM-generated judgments. For cold-start and resource-limited scenarios, the distilled variant FACTGUARD-D is optimized for efficiency and resources.Experiment results on Weibo21 and GossipCop datasets show that FACTGUARD outperforms baselines with each module proven effective by ablation studies, while FACTGUARD-D achieves a strong balance between accuracy and speed.

Future Work. Future directions include: (1) developing customized methods for Chinese and English fake news; (2) optimizing the model for edge deployment; (3) enhancing interpretability of usability evaluation module to improve transparency and credibility; (4) exploring the role of text style at different stages of news dissemination detection; and (5) considering the benchmark data contamination of the employed LLMs, and extending cross-domain adaptation across emerging platforms and multimodal signals.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grants 62162067 and 82360280, in part by the Yunnan Province Special Project under Grant 202403AP140021, in part by the Yunnan Fundamental Research Project under Grant 202401AT070474, in part by the Yunnan University Interdisciplinary Research Team under Grant CZ22621802, and in part by Scientific Research and Innovation Project of Postgraduate Students in the Academic Degree of Yunnan University under KC-252512080.

References

Ajao, O.; Bhowmik, D.; and Zargari, S. 2019. Sentiment aware fake news detection on online social networks. In *ICASSP*, 2507–2511.

Amado, B. G.; Arce, R.; and Fariña, F. 2015. Undeutsch hypothesis and Criteria Based Content Analysis: A meta-analytic review. *The European Journal of Psychology Applied to Legal Context*, 7: 3–12.

Avaaz. 2020. Facebook's algorithm: A major threat to public health. https://secure.avaaz.org/campaign/en/facebook_threat_health/. Accessed: 2025-07-22.

Bursztyn, L.; Rao, A.; Roth, C. P.; and Yanagizawa-Drott, D. H. 2020. Misinformation During a Pandemic. Technical Report 27417, National Bureau of Economic Research.

Castillo, C.; Mendoza, M.; and Poblete, B. 2011. Information credibility on twitter. In *WWW*, 675–684.

Chan, C.-M.; Chen, W.; Su, Y.; Yu, J.; Xue, W.; Zhang, S.; Fu, J.; and Liu, Z. 2024. ChatEval: Towards Better LLM-based Evaluators through Multi-Agent Debate. In *ICLR*.

Chen, X.; Huang, X.; Gao, Q.; Huang, L.; and Liu, G. 2025. Enhancing text-centric fake news detection via external knowledge distillation from LLMs. *Neural Networks*, 187: 107377.

Cui, J.; Kim, K.; Na, S. H.; and Shin, S. 2022. Meta-path-based fake news detection leveraging multi-level social context information. In *CIKM*, 325–334.

DeepSeek-AI. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. arXiv:2501.12948.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, 4171–4186.

Galtung, J.; and Ruge, M. H. 1965. The structure of foreign news: The presentation of the Congo, Cuba and Cyprus crises in four Norwegian newspapers. *Journal of peace research*, 2: 64–90.

Giachanou, A.; Rosso, P.; and Crestani, F. 2019. Leveraging emotional signals for credibility detection. In *SIGIR*, 877–880.

Granik, M.; and Mesyura, V. 2017. Fake news detection using naive Bayes classifier. In *IEEE UkrCon*, 900–903.

Grinberg, N.; Joseph, K.; Friedland, L.; Swire-Thompson, B.; and Lazer, D. 2019. Fake news on Twitter during the 2016 US presidential election. *Science*, 363: 374–378.

Hinton, G.; Vinyals, O.; and Dean, J. 2015. Distilling the knowledge in a neural network. arXiv:1503.02531.

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 *AAAI*, 22105–22113.

Hu, X.; Guo, Z.; Wu, G.; Liu, A.; Wen, L.; and Yu, P. S. 2022. CHEF: A pilot Chinese dataset for evidence-based fact-checking. In *NAACL*, 3362–3376.

Kim, S.; Kim, D.; Park, C.; Lee, W.; Song, W.; Kim, Y.; Kim, H.; Kim, Y.; Lee, H.; Kim, J.; et al. 2024. SOLAR 10.7 B: Scaling Large Language Models with Simple yet Effective Depth Up-Scaling. In *NAACL*, 23–35.

Liu, Y.; Liu, Y.; Zhang, X.; Chen, X.; and Yan, R. 2025. The truth becomes clearer through debate! Multi-agent systems with large language models unmask fake news. In *SIGIR*, 504–514.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. Roberta: A robustly optimized bert pretraining approach. arXiv:1907.11692.

Mu, Y.; Bontcheva, K.; and Aletras, N. 2023. It's about Time: Rethinking Evaluation on Rumor Detection Benchmarks using Chronological Splits. In *EACL*, 736–743.

Nan, Q.; Cao, J.; Zhu, Y.; Wang, Y.; and Li, J. 2021. MDFEND: Multi-domain Fake News Detection. In *CIKM*, 3343–3347.

Nan, Q.; Sheng, Q.; Cao, J.; Hu, B.; Wang, D.; and Li, J. 2024. Let silence speak: Enhancing fake news detection with generated comments from large language models. In *CIKM*, 1732–1742.

OpenAI. 2023. GPT-3.5 Turbo Model. https://platform.openai.com/docs/models/gpt-3.5-turbo. Accessed: 2025-07-22.

OpenAI. 2024. GPT-4o mini: Advancing Cost-Efficient Intelligence. https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/. Accessed: 2025-07-22.

Potthast, M.; Kiesel, J.; Reinartz, K.; Bevendorff, J.; and Stein, B. 2018. A Stylometric Inquiry into Hyperpartisan and Fake News. In *ACL*, 231–240. Melbourne, Australia: Association for Computational Linguistics.

Przybyla, P. 2020. Capturing the style of fake news. In *AAAI*, 490–497. New York, USA: AAAI Press.

Qian, F.; Gong, C.; Sharma, K.; and Liu, Y. 2018. Neural User Response Generator: Fake News Detection with Collective User Intelligence. In *IJCAI*, 3834–3840.

Rashkin, H.; Choi, E.; Jang, J. Y.; Volkova, S.; and Choi, Y. 2017. Truth of Varying Shades: Analyzing Language in Fake News and Political Fact-Checking. In *EMNLP*, 2931–2937. Copenhagen, Denmark: Association for Computational Linguistics.

Refutation, W. R. 2021. Weibo Rumor Refutation Report 2021. https://weibo.com/1866405545/LcFuud7ml#repost. Accessed: 2025-07-22.

Sheng, Q.; Cao, J.; Zhang, X.; Li, R.; Wang, D.; and Zhu, Y. 2022. Zoom Out and Observe: News Environment Perception for Fake News Detection. In *ACL*, 4543–4556.

Shu, K.; Cui, L.; Wang, S.; Lee, D.; and Liu, H. 2019. defend: Explainable fake news detection. In *KDD*, 395–405.

Shu, K.; Mahudeswaran, D.; Wang, S.; Lee, D.; and Liu, H. 2020. Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. *Big data*, 8: 171–188.

Wang, L. Z.; Ma, Y.; Gao, R.; Guo, B.; Zhu, H.; Fan, W.; Lu, Z.; and Ng, K. C. 2024. Megafake: a theory-driven dataset of fake news generated by large language models. arXiv:2408.11871.

Wang, W. Y. 2017. "Liar, Liar Pants on Fire": A New Benchmark Dataset for Fake News Detection. In *ACL*, 422–426.

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 *KDD*, 849–857.

Wu, J.; Guo, J.; and Hooi, B. 2024. Fake News in Sheep's Clothing: Robust Fake News Detection Against LLM-Empowered Style Attacks. In *KDD*, 3367–3378.

Xu, Z.; Jain, S.; and Kankanhalli, M. 2024. Hallucination is inevitable: An innate limitation of large language models. arXiv:2401.11817.

Zhang, X.; Cao, J.; Li, X.; Sheng, Q.; Zhong, L.; and Shu, K. 2021. Mining dual emotion for fake news detection. In *WWW*, 3465–3476.

Zhang, Z.; Liu, Q.; Hu, Z.; Zhan, Y.; Huang, Z.; Gao, W.; and Mao, Q. 2024. Enhancing Fairness in Meta-learned User Modeling via Adaptive Sampling. In *WWW*, 3241–3252. Singapore: ACM.

Zhong, Q.; Ding, L.; Liu, J.; Du, B.; and Tao, D. 2023. Can chatgpt understand too? a comparative study on chatgpt and fine-tuned bert. arXiv:2302.10198.

Zhu, Y.; Sheng, Q.; Cao, J.; Li, S.; Wang, D.; and Zhuang, F. 2022. Generalizing to the future: Mitigating entity bias in fake news detection. In *SIGIR*, 2120–2125.