

CogniDir: Combating Cognitive Malicious Comments via Adaptive Distributional Learning for Robust Fake News Detection

Zhao Tong^{1,2*}, Chunlin Gong^{3*}, Yimeng Gu⁴, Haichao Shi¹, Qiang Liu⁵,
Shu Wu^{5†}, Xingcheng Xu⁶, Xiao-Yu Zhang^{1†}

¹Institute of Information Engineering, Chinese Academy of Sciences

²School of Cyber Security, University of Chinese Academy of Sciences

³University of Minnesota, ⁴Queen Mary University of London

⁵New Laboratory of Pattern Recognition (NLPR),

State Key Laboratory of Multimodal Artificial Intelligence Systems (MAIS),
Institute of Automation, Chinese Academy of Sciences

⁶Shanghai AI Laboratory

tongzhao@iie.ac.cn, gong0226@umn.edu

shu.wu@nlpr.ia.ac.cn, zhangxiaoyu@iie.ac.cn

 [Project Page](#) |  [GitHub Code](#)

Abstract

The proliferation of Large Language Models (LLMs) has enabled a new class of psychologically grounded malicious comments, shifting fake news attacks from surface-level textual noise to deep cognitive and logical manipulation. This shift severely undermines existing detectors, which conventionally rely on static attack assumptions and fixed training distributions. To bridge this gap, we introduce CogniDir, an adaptive distributional learning framework that reformulates robust detection as a dynamic data mixture optimization problem for social media content safety. Grounded in cognitive psychology, we first formalize mechanism-specific cognitive adversarial paradigms to systematically expose deep-seated detector vulnerabilities. To address the vulnerability heterogeneity, CogniDir derives an information-theoretic score coupling empirical accuracy with probabilistic confidence, which is then mapped to adaptive sampling proportions through a Dirichlet-mean parameterization. This formulation enables smooth, feedback-driven reallocation of training exposure toward the most brittle attack mechanisms. Experimental results on three benchmarks demonstrate that CogniDir yields state-of-the-art robustness, improving F1 scores by up to 17.9% over competitive baselines under heterogeneous, AI-generated adversarial pressures. Our codes are available at [this website](#)

1 Introduction

The paradigm of fake news detection has been fundamentally challenged by the rapid proliferation of Large Language Models (LLMs)(Tong

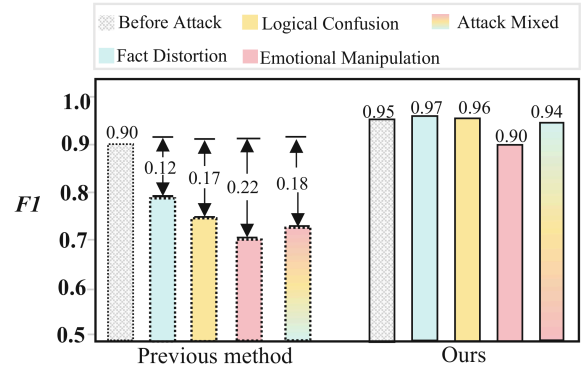


Figure 1: Group-wise F1 comparison on Weibo16(Ma et al., 2016). Previous method (Shu et al., 2019) typically suffers a performance drop of more than 18% under different types of malicious comments. Yet, our model is able to significantly narrow this performance gap, achieving 0.94 in F1 across diverse malicious comments.

et al., 2026b,a). Beyond generating standard fake news, malicious attackers can now deploy automated, LLM-augmented user comments that act as sophisticated adversarial perturbations against existing detectors (Chen and Shu, 2024; Nutsugah et al., 2025). Unlike conventional text perturbations that rely on surface-level lexicon alterations, these LLM-generated malicious comments systematically exploit cognitive and algorithmic blind spots through strategic logical manipulation and psychological framing (Tong et al., 2025; Hu et al., 2025). Consequently, they induce severe alignment degradation and classification blindness in state-of-the-art detection systems. Maintaining proactive robustness under such heterogeneous, mechanism-driven adversarial pressures has thus emerged as a critical yet under-explored frontier in digital content security.

*These authors contributed to the work equally.

†To whom correspondence should be addressed.

To counter these evolving threats, the current detectors leverage crowd intelligence to fortify verification boundaries (Guo et al., 2022; Nan et al., 2025). However, they inevitably suffer from severe *distributional mismatch* under realistic deployment, driven by a continuous causal failure chain: (i) **Mechanism Homogenization**: Existing frameworks conventionally treat diverse cognitive malicious comments as generic text noise, making them fundamentally incapable of distinguishing fine-grained interference pathways like fact distortion, logical confusion, and emotional manipulation (see Fig. 1). (ii) **Static Data Assumptions**: Because they fail to decouple these distinct cognitive mechanisms, their subsequent defenses rely entirely on fixed, homogeneous training distributions that cannot track the fluid, evolutionary nature of adaptive LLM strategies. (iii) **Uniform Training Dynamics**: Crucially, this reliance on invariant data forces the optimization to apply static data mixing across all attack classes, implicitly assuming identical defense difficulty. Consequently, the optimization trajectory rapidly overfits to shallow textual patterns, leaving group-wise vulnerability heterogeneity unaddressed and causing robustness to plateau under novel attack distributions.

To bridge this gap, we propose **CogniDir**, a framework that reframes robust detection as a *dynamic distributional optimization problem*. Grounded in cognitive psychology (Pennycook and Rand, 2021; Park et al., 2025a; Moradi and Samwald, 2021; Yeh et al., 2024; Liu et al., 2024), CogniDir addresses two core challenges: First, to overcome the scarcity of mechanism-labeled adversarial data, we design a *cognitive-grounded synthesis pipeline* that leverages LLMs to generate diverse, category-specific perturbations, establishing a scalable foundation for AI safety research. Second, to achieve adaptive robustness, we formalize training exposure allocation as an optimization problem over the probability simplex. By deriving an information-theoretic vulnerability metric that couples accuracy with probabilistic confidence, we map group-wise brittleness to adaptive sampling proportions through a Dirichlet-mean parameterization, enabling smooth, feedback-driven reallocation of sampling budgets. This transforms adversarial training into a self-correcting curriculum that continuously targets the model’s weakest mechanisms while ensuring cross-lingual transferability.

Our contributions are summarized as follows:

- We introduce a cognitive-grounded adversarial synthesis pipeline to generate mechanism-labeled malicious comments, which improves training data diversity and fidelity.
- We propose CogniDir, a distributional optimization framework in which an information-theoretic vulnerability metric adaptively maps group-wise vulnerability for feedback-driven allocation.
- We demonstrate that CogniDir achieves state-of-the-art robustness across three cross-lingual benchmarks, narrowing the vulnerability gap under heterogeneous attacks while maintaining high baseline accuracy.

2 Related Work

2.1 Comment-based Fake News Detection

News comments provide crowd intelligence that complements news content for fake news detection (Shu et al., 2019; Yang et al., 2023a; Nan et al., 2024). Early methods co-model news and comments via attention mechanisms to jointly capture check-worthy content (Shu et al., 2019), sequential features (Yang et al., 2023a), or crucial comment-context interactions (Wang et al., 2023). However, diverse comments are not always available. To address this, recent studies explore data augmentation: (Nan et al., 2024) use LLMs to generate alternative comments, (Wan et al., 2024) simulate user-news interactions with generated reactions, and (Yanagi et al., 2020) augment sets by generating new comments. Meanwhile, other works deepen comment modeling, utilizing comments to guide text–image cross-modal attention (Yi et al., 2025) or modeling interactions between news emotion and comment semantics (Tan and Zhang, 2026).

Despite these advances, the vulnerability of comment-based detectors to malicious comments remains understudied. In this work, we systematically evaluate and attack detectors through three categories of comment perturbations, and propose a novel framework to defend against such attacks.

2.2 Attack on Fake News Detection

Despite advances in fake news detection, attacking these detectors remains underexplored (Le et al., 2020; Zhu et al., 2024; Park et al., 2025b). Existing attacks primarily target news content, propagation structures, or comments. Content- and

structure-based attacks modify writing styles (Wu et al., 2024), apply meaning-preserving transformations (Przybyła et al., 2024), or perturb propagation networks (Zhu et al., 2024) and user graphs (Wang et al., 2024b). In contrast, comment-based attacks manipulate user responses while keeping the news unchanged. Specifically, attackers generate malicious comments (Le et al., 2020) or transfer real comments from other articles (Koren et al., 2025). Consequently, adversarial training has been explored to improve robustness (Park et al., 2025b; Sun et al., 2025).

Content and comment attacks correspond to distinct attack surfaces and capabilities: modifying publisher-controlled text versus injecting comments in open environments. Focusing on the latter, our work investigates heterogeneous cognitive mechanisms in malicious comments and enhances comment-aware detectors via vulnerability-guided adaptive adversarial training.

3 Problem Definition

Comment-based fake news detection predicts the authenticity of a given news based on the news content and its associated comments. Formally, let the news dataset be $D = \{(x_i, C_i, y_i)\}_{i=1}^N$, where x_i denotes the content of the i -th news article, $C_i = \{C_i^j\}_{j=1}^M$ denotes its associated original comment set, and $y_i \in \{0, 1\}$ is the news authenticity label (1 for fake and 0 for real). Given (x_i, C_i) , the goal is to learn a detector f that predicts y_i .

During the attack, the news content x_i remains unchanged, while malicious comments C_i^{adv} are injected into the original comment set, forming $C_i' = C_i \cup C_i^{adv}$. For an originally correct prediction $f(x_i, C_i) = y_i$, an attack succeeds when $f(x_i, C_i') \neq y_i$. We consider untargeted attacks that include both real-to-fake and fake-to-real prediction flips (see Appendix A).

4 Methodology

In this section, we present **CogniDir**, a group-adaptive framework for robust fake news detection under malicious comment attacks. As illustrated in Fig. 2, CogniDir first constructs mechanism-specific malicious comments with LLMs, covering fact distortion, logical confusion, and emotional manipulation. It then uses InfoDirichlet Resampling to estimate group-wise vulnerabilities after each training epoch and adaptively reallocate the sampling proportions of different mali-

cious comment types. Finally, the trained detector is evaluated under mixed adversarial comment settings to assess its robustness against heterogeneous attack mechanisms.

4.1 Multi-Type Malicious Comments Generation

To better approximate the characteristics of malicious comments in real-world social environments, we categorize attackers’ motivations into three complementary mechanisms based on cognitive psychology (Tversky and Kahneman, 1981, 1978; Tajfel et al., 1971): fact distortion, logical confusion, and emotional manipulation. However, existing public datasets provide only task-level annotations, such as news veracity or reply stance, without fine-grained annotations of cognitive manipulation mechanisms. Specifically, Weibo16 and Weibo20 provide news-level veracity labels, while RumourEval-19 additionally provides reply-level stance labels. None of these datasets contains cognitive-mechanism annotations, as detailed in Appendix B.

To address this gap, we use the strong reasoning capabilities of LLM to automatically generate malicious comments aligned with these mechanisms, yielding a dataset with explicit mechanism labels. Concretely, we design a prompt to simulate how real users produce malicious comments given a news article, while preserving linguistic naturalness and task relevance (see Appendix C for dataset quality evaluation). Moreover, to reduce generator-specific style bias, we generate attacks using three models of different scales (Mistral-7B (Jiang et al., 2024), Gemma-2B (Team et al., 2024), and Qwen-32B (Hui et al., 2024)), and then perform stratified random sampling and combine them to form the final malicious comment dataset. We further examine potential generator-specific shortcuts by comparing single and mixed-generator settings, with the results reported in Appendix D. In summary, we design and generate a high-quality malicious comment dataset with diverse writing styles, enabling the robustness evaluation of malicious comment detection models against potential real-world malicious comments (Goodfellow et al., 2015).

Prompt Design. As shown in Fig. 3, we adopt an *understand-first, then-generate* CoT prompt to synthesize realistic malicious comments. By first identifying salient content cues and then conditioning generation on them, the prompt improves

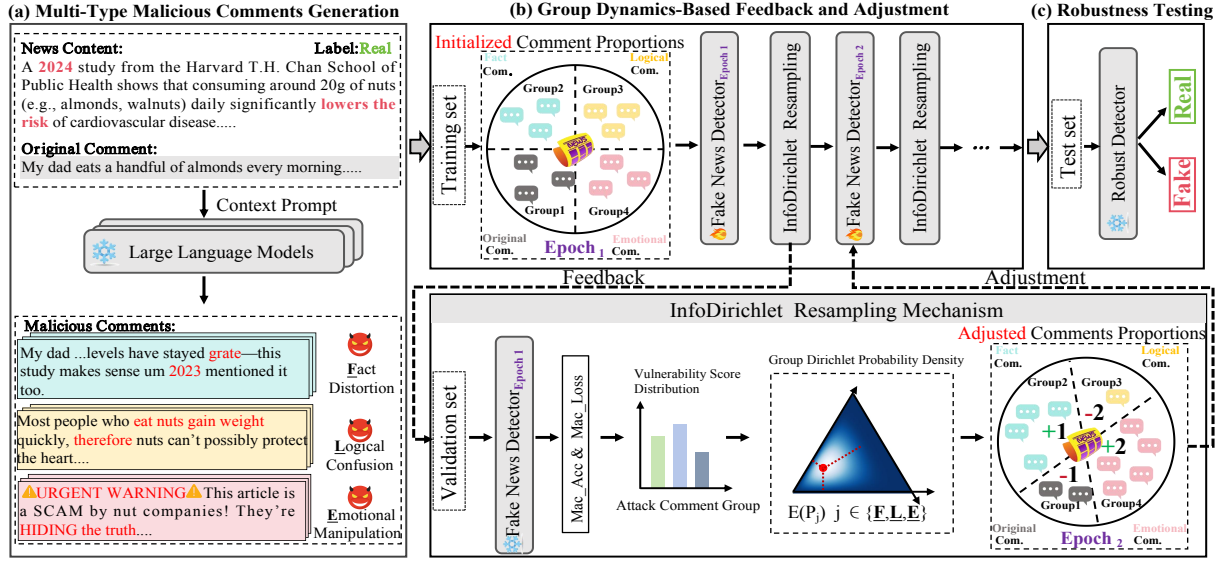


Figure 2: Overview of **CogniDir**. The framework consists of three stages: mechanism-specific malicious comment generation, feedback-driven sampling adjustment via InfoDirichlet Resampling, and robustness evaluation under mixed adversarial comment settings.

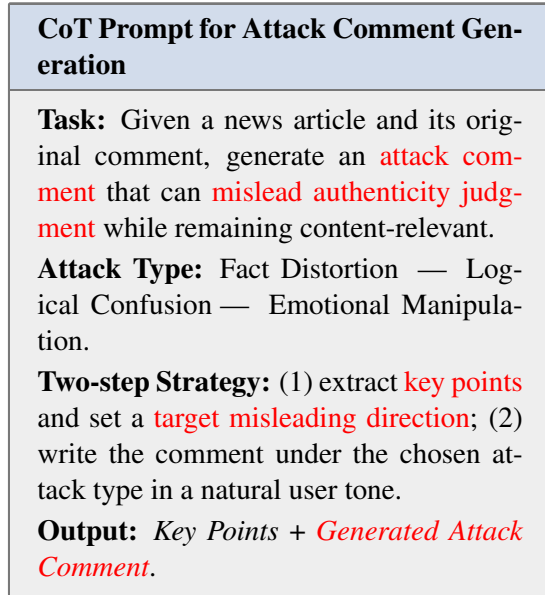


Figure 3: Two-step CoT prompting strategy for generating attack comments with a specified attack type: understand the content first, then produce a realistic, content-grounded misleading comment.

content grounding and reduces off-topic or implausible attacks. Meanwhile, specifying the attack type and requiring a natural user tone enables controllable generation across the three attack mechanisms. To assess the realism and mechanism fidelity of the generated comments, we conduct a human evaluation on relevance, naturalness, attack-type consistency, and misleading potential, as shown in Appendix E. Further, to examine their

real-world relevance, we qualitatively compare the generated attacks with real-world misleading user content exhibiting corresponding cognitive mechanisms, as shown in Appendix F.

Dataset setting. To support group-adaptive resampling, we organize the training, validation, and test sets according to attack mechanisms rather than only for static offline evaluation. Specifically, we aggregate all comment types, paired with their source news articles, into a unified training pool. Given a fixed budget of M comments per instance, IDR dynamically controls the sampling proportions of different malicious groups from this pool, and we set $M = 6$ according to the available generated comments. During training, group-specific validation sets, each containing comments from a single attack mechanism, are used to estimate attack-wise vulnerability and guide subsequent sampling updates. For testing, we uniformly sample from all malicious groups and combine them into mixed adversarial comment sets, enabling evaluation under multiple coexisting attack mechanisms.

4.2 Group Dynamics-Based Feedback and Adjustment

After getting a malicious comments combined datasets, we further need a robust detector to deal with the coexistence of multiple attack mechanisms. However, conventional methods typically train on all comments jointly without explicitly

characterizing the detector’s differential vulnerabilities across attack types. As a result, training is dominated by an averaged objective, which often leads to imbalanced robustness across groups.

To address this, we introduce *Group Dynamics-Based Feedback and Adjustment* in CogniDir, a feedback-driven training module that consists of an *InfoDirichlet Resampling* mechanism to compute group-wise vulnerability scores and adjust the proportions, thereby increasing the exposure of specific vulnerable group. By iterating *Adaptive Resampling Training Loop*, the model can consistently focus on high vulnerability groups, gradually improving and balancing robustness across all groups.

4.2.1 Detector Training

At epoch t , we construct the comment set $\mathcal{C}_i^{(t)}$ for each news instance according to the current group-wise sampling distribution. The first epoch uses a uniform allocation over all attack groups, while subsequent epochs update this distribution based on the vulnerability scores estimated from validation feedback. Given the sampled training data, we optimize a shared detector f_θ by minimizing the empirical loss:

$$\theta^* = \arg \min_{\theta} \sum_{i=1}^N \mathcal{L}(f_\theta(x_i, \mathcal{C}_i^{(t)}), y_i). \quad (1)$$

Upon completing this training round, the model is specifically enhanced on the previous updated group distribution, thereby progressively improve robustness imbalance across different groups.

4.2.2 InfoDirichlet Resampling

To further assess the model’s robustness across different groups after this training round, we evaluate its vulnerability to different attack groups and adjust the training data distribution accordingly. Therefore, this section focuses on two key aspects: (i) how to quantify group-wise vulnerability from validation signals; and (ii) how to convert these vulnerability measures into meaningful sampling proportions defined by a group probability distribution, enabling adaptive allocation of training resources across groups.

(i) Obtain Vulnerability Score. After each training round, we evaluate the current detector on group-specific validation sets to get their performance. Notably, when faced with malicious comments, the detector needs to classify the arti-

cle with multiple potentially contradictory comments, which often results in uncertain prediction. Since accuracy alone cannot fully capture this phenomenon, we introduce probabilistic confidence as an additional metric to better characterize the model’s robustness against specific types of attacks (Guo et al., 2017).

Concretely, we adopt an information-theoretic (Cover and Thomas, 1991) view by focusing on the *correctness event*. Let $Z = [\hat{y} = y] \in \{0, 1\}$ represent whether a prediction is correct. For each attack group $j \in \{F, L, E\}$, we evaluate Detector on the corresponding validation set val_j (consisting only of samples from group j) and record the average cross-entropy loss and accuracy:

$$m_{j,\text{loss}}, m_{j,\text{acc}} = \text{Detector}(val_j), a_j \triangleq m_{j,\text{acc}}. \quad (2)$$

Here, a_j is the empirical frequency of the correctness event Z . We further convert the average negative log-likelihood into an equivalent probability mass,

$$q_j \triangleq \exp(-m_{j,\text{loss}}), \quad (3)$$

which corresponds to the geometric-mean probability assigned to the true label (derivation in the Appendix G).

Based on the above, we define the group-wise vulnerability score, s_j , as the Bernoulli cross-entropy from the empirical parameter a_j to the model-implied parameter q_j :

$$s_j \triangleq -a_j \log q_j - (1 - a_j) \log(1 - q_j). \quad (4)$$

This admits the standard information-theoretic decomposition:

$$s_j = H(\text{Bern}(a_j)) + D_{\text{KL}}(\text{Bern}(a_j) \parallel \text{Bern}(q_j)), \quad (5)$$

where the entropy term depends only on empirical correctness, while the KL divergence quantifies the mismatch between empirical correctness and the model-implied probabilistic characterization.

In short, this information-theoretic coupling of accuracy and confidence yields a group-wise vulnerability score s_j for each attack group, which is subsequently used to derive the sampling/allocation probabilities in the next stage.

(ii) Adjust Comments Proportions. After we obtain the vulnerability score $\{s_F, s_L, s_E\}$, we

convert them into a group-level probability distribution to control the sampling budget across adversarial groups in subsequent training rounds. We aim to construct an allocation vector $p = (p_F, p_L, p_E)$ satisfying: (i) p lies on the probability simplex (i.e., $p_j \geq 0$ and $\sum_j p_j = 1$), and (ii) p varies smoothly with the relative magnitudes of $\{s_j\}$, enabling adaptive updates of the sampling strategy throughout training.

To obtain a group-probability mapping defined on the probability simplex, we model p using a Dirichlet distribution (Bishop, 2006):

$$p \sim \text{Dir}(\alpha), \quad f(p; \alpha) = \frac{1}{B(\alpha)} \prod_{j \in \{F, L, E\}} p_j^{\alpha_j - 1}, \quad (6)$$

where $B(\alpha) = \frac{\prod_j \Gamma(\alpha_j)}{\Gamma(\sum_j \alpha_j)}$, and $\alpha = (\alpha_F, \alpha_L, \alpha_E)$ denotes the concentration parameters. Here, α serves as a bridge from vulnerability scores to a probability distribution: larger α_j leads to higher expected mass for group j , while the normalization inherent in the Dirichlet family ensures $\sum_j p_j = 1$.

We define α via a simple linear mapping from the vulnerability scores:

$$\alpha_j \triangleq s_j + 1, \quad j \in \{F, L, E\}, \quad (7)$$

where the constant offset guarantees $\alpha_j > 0$, ensuring parameter validity. We adopt the mean of the Dirichlet distribution as a deterministic allocation strategy (see Appendix H), which avoids the variance of random sampling and allows continuous updates with respect to $\{s_j\}$:

$$\bar{p}_j \triangleq \mathbb{E}[p_j] = \frac{\alpha_j}{\sum_{k \in \{F, L, E\}} \alpha_k}. \quad (8)$$

This yields a smooth probability distribution $\bar{p}$ over the simplex, which is used to allocate the budget for malicious comments. Specifically, let M denote the total number of comment set assignable to the three adversarial groups per instance. We allocate new comment set as follows:

$$\mathcal{C}_i^{(t+1)} = \bigcup_{j \in \{F, L, E\}} \text{Sample}(\lfloor \bar{p}_j M \rfloor). \quad (9)$$

In summary, Dirichlet Expectation Allocation provides a principled mechanism that smoothly maps vulnerability scores to group-level probability distributions, enabling stable and well-structured control over group-wise sampling

throughout training. An equivalent entropy-regularized optimization interpretation and the properties of this allocation are provided in Appendix H.

(iii) Adaptive Resampling Training Loop. After applying InfoDirichlet Resampling to obtain a newly sampled comment set $\mathcal{C}_i^{(t+1)}$, we proceed to the next $t + 1$ epoch of Detector Training. In summary, this resampling step closes the feedback loop by reallocating training exposure toward the most vulnerable groups, enabling targeted improvement and more balanced robustness across attack types.

4.3 Robustness Testing

After the model converges at epoch T , we apply the detector f_{θ^*} for classification. Given a test instance $(x_i, \mathcal{C}_i)$, we compute the prediction as:

$$\hat{y}_i = f_{\theta^*}(x_i, \mathcal{C}_i'), \quad \hat{y}_i \in \{0, 1\}. \quad (10)$$

In conclusion, by iteratively adjusting group-wise sampling proportions during training, we progressively reinforce the detector along its weakest attack groups and obtain a final model f_{θ^*} that is robust and well-balanced across all malicious comment groups.

5 Experiment

In this section, we conduct comprehensive experiments to answer the following research questions:

- **RQ1: Attack Effectiveness.** How do malicious comments affect fake news detectors?
- **RQ2: Robustness Enhancement.** Can Cog-niDir improve robustness against malicious comments?
- **RQ3: Adaptive Allocation.** Does adaptive allocation reduce vulnerability gaps, and which components matter?

5.1 Dataset

Considering for cultural and linguistic variation in information interpretation (Nisbett et al., 2001), we evaluate our model on the Chinese datasets Weibo16 (Ma et al., 2016) and Weibo20 (Rao et al., 2021), and the English dataset RumourEval-19 (Gorrell et al., 2019). This cross-lingual setting enables a robust assessment under diverse social and discourse environments. Dataset statistics are summarized in Tab. 1.

Table 1: Dataset statistics. pcs: number of news pieces; com: number of comments.

Split	Veracity	Weibo-16		Weibo-20		RumourEval-19	
		pcs	com	pcs	com	pcs	com
Training	Fake	369	4,428	702	8,424	142	1,704
	Real	676	8,112	987	11,844	214	2,568
	Total	1,045	12,540	1,689	20,268	356	4,272
Validation	Fake	154	924	318	1,908	50	600
	Real	316	1,896	421	2,526	82	984
	Total	470	2,820	739	4,434	132	1,584
Test	Fake	70	420	119	714	16	96
	Real	112	672	177	1,062	39	234
	Total	182	1,092	296	1,776	55	330
All	Fake	593	5,772	1,139	11,046	208	2,400
	Real	1,104	10,680	1,585	15,432	335	3,786
	Total	1,697	16,452	2,724	26,478	543	6,186

5.2 Implementation Details

Model Architecture. For the Detector, we adopt a comment-aware classifier. We use a shared pre-trained BERT encoder to extract representations for the news article and each associated comment. These representations are refined by a self-attention layer and be concatenated; finally, an MLP classifier outputs the binary classification result.

Attack Settings. To systematically assess robustness under diverse adversarial behaviors, we construct two complementary evaluation settings: *group-specific* and *mixed*. Group-specific sets are formed by composing each test instance exclusively with comments from a single attack type, enabling targeted diagnosis of attack-wise failure modes and worst-case susceptibility. In contrast, the mixed set integrates multiple attack types within each instance to approximate realistic comment streams and measure robustness under combined adversarial interference. In conclusion, these two settings jointly capture both per-attack brittleness and practical robustness under real-world attack mixtures. Throughout evaluation, O denotes performance on the original comment set, A denotes performance on the attacked comment set before robust training, and R denotes performance on the same attacked test inputs after robust training.

5.3 Baselines

We select representative baselines from two categories: (i) **LLM-Only**: We use four LLMs of different sizes: Gemma-2-2B (Team et al., 2024), Mistral-7B (Jiang et al., 2023), Llama-3-8B (Grattafiori et al., 2024), and Qwen2.5-

32B (Hui et al., 2024)). (ii) **Deep-Learning Based**: we adopt five representative deep-learning fake news detectors DEFEND (Shu et al., 2019), Dual-CAN (Yang et al., 2023b), GenFEND (Nan et al., 2024), L-Defense (Wang et al., 2024a), ARG (Hu et al., 2024), and our proposed **CogniDir**. Details of the baselines and evaluation protocols are provided in the Appendix I.

Robustness Testing Analysis. CogniDir exhibits robustness under malicious comments, mitigating performance degradation caused by malicious inputs. Traditional detectors suffer an average performance drop of over 18% under different attack types, whereas CogniDir narrows this gap, maintaining an F1 above 0.94 across attack categories. This robustness stems from our adaptive training strategy, which dynamically re-allocates learning focus toward more vulnerable attack groups via the InfoDirichlet Resampling. To assess result stability, we additionally repeat CogniDir and GenFEND with three matched random seeds. Detailed mean±standard deviation and paired significance tests are reported in Appendix J.

LLM-Only vs. Deep-Learning Based Analysis. When comparing LLM-only and Deep-Learning Based detectors, we observe that LLMs such as Qwen2.5-32B perform competitively in non-adversarial settings but exhibit pronounced vulnerability under attack, e.g., F1 on Weibo16 drops sharply from 0.831 to 0.282. In contrast, deep-learning models like DEFEND and GenFEND show stronger baseline robustness yet still leave considerable room for improvement. CogniDir bridges this gap by integrating a deep-learning architecture with LLM-generated adversarial examples and adaptive training. It leverages LLMs to synthesize diverse and realistic attack samples, while employing a dynamic training mechanism to harden the detector. This combined strategy allows CogniDir to achieve both high initial accuracy and sustained robustness, outperforming all LLM-only and deep-learning baselines in adversarial environments.

RQ1-RQ2. Malicious comments consistently degrade existing detectors, while CogniDir outperforms the strongest baselines by **8.9–21.1%** under attack.

5.4 Robust Analysis

We evaluate the robustness of CogniDir under different malicious comment numbers on Weibo16,

Table 2: Performance comparison between **CogniDir** and baselines on Weibo16, Weibo20, and RumourEval-19 datasets. "O" denotes original detection performance, "A" denotes performance following adversarial attack, and "R" denotes performance after robust training (with improvements over the attacked setting)—all reported metrics herein are F1 scores. The best performance is highlighted in bold. $\Delta Improve$ represents our method's relative improvements over the second best performance in percentage.

Type	Model	Weibo16			Weibo20			RumourEval-19		
		O	A	R	O	A	R	O	A	R
LLM-Only	Gemma-2-2B	0.322	0.291	0.274	0.345	0.292	0.326	0.294	0.238	0.275
	Mistral-7B	0.686	0.334	0.495	0.654	0.315	0.375	0.468	0.184	0.252
	Llama-3-8B	0.598	0.324	0.408	0.582	0.363	0.408	0.377	0.219	0.276
	Qwen2.5-32B	0.831	0.282	0.716	0.823	0.367	0.541	0.553	0.343	0.412
Deep-Learning Based	dEFEND	0.902	0.727	0.745	0.884	0.557	0.585	0.648	0.407	0.687
	Dual-CAN	0.895	0.366	0.578	0.872	0.436	0.580	0.628	0.377	0.546
	GenFEND	0.915	0.578	0.867	0.892	0.640	0.761	0.662	0.456	0.606
	L-Defense	0.882	0.706	0.759	0.860	0.666	0.702	0.610	0.343	0.709
	ARG	0.884	0.619	0.820	0.812	0.576	0.803	0.712	0.440	0.609
	CogniDir	0.955	0.792	0.945	0.936	0.728	0.947	0.822	0.552	0.812
$\Delta Improve$		+4.4%	+8.9%	+9.0%	+4.9%	+9.3%	+17.9%	+15.5%	+21.1%	+14.5%

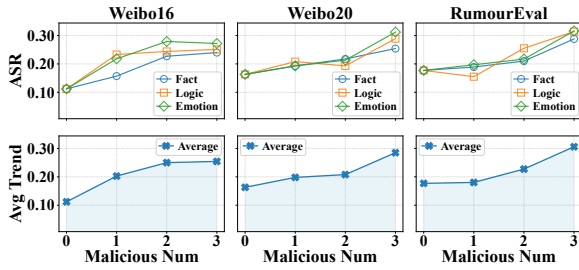


Figure 4: ASR **before** robust training under varying attack types and numbers. Top: Fact/Logic/Emotion ASR under the number of injected malicious comments across three datasets. Bottom: mean ASR averaged over attack types at each attack number.

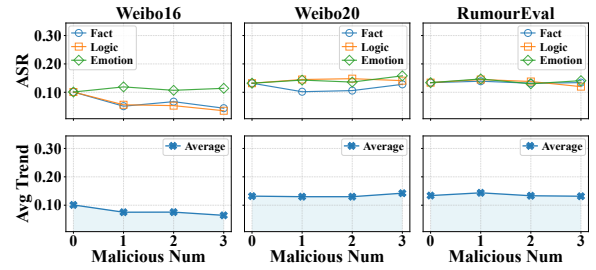


Figure 5: ASR **after** robust training under varying attack types and numbers. Top: Fact/Logic/Emotion ASR under the number of injected malicious comments across three datasets. Bottom: mean ASR averaged over attack types at each attack number.

Weibo20, and RumourEval-19. We first measure attack success rate(ASR) without adversarial training (as shown Fig. 4), then assess improvements after incorporating adversarial samples during training (as shown Fig. 5), analyzing how robustness scales with attack intensity across categories.

Effect of malicious comment numbers. When increasing the number of malicious comments from 0 to 3 in Fig.4, the ASR across datasets rises from 0.145 to 0.287, an increase of nearly 98%. This trend confirms that denser adversarial inputs accumulate semantic noise and contradictory cues, which destabilize the model’s feature representation and decision boundary. Hence, higher attack intensity amplifies uncertainty and weakens detection reliability across all attack types.

Effect of Malicious Comment Types. As shown in Fig. 4, the average Attack Success Rates (ASR) are 0.206 (fact distortion), 0.241 (logical confusion), and 0.283 (emotional manipulation). This

indicates the model is most robust to fact distortion, whose lexical noise is contextually mitigated; moderately affected by logical confusion; and most vulnerable to emotional manipulation, which alters sentiment to severely mislead authenticity judgment.

Effect of Robust Training. As shown in Fig.5, after incorporating adversarial group training, the mean ASR drops to 0.116, representing an overall 45% reduction compared with the untrained setting. Moreover, inter-attack variance decreases sharply (from over 0.07 to below 0.03), indicating that the model becomes both more accurate and more stable across attack types and densities. These findings suggest that our robust training strategy enables the model to learn generalized and invariant representations, maintaining consistent detection performance even under complex adversarial scenarios.

RQ2. CogniDir remains robust across attack types and different numbers of malicious comments, reducing the mean ASR by **45%**.

5.5 Convergence Analysis

To empirically verify the adaptive robustness improvement across different attack categories during the training process, we continuously track the validation accuracy on three distinct adversarial groups: Fact Distortion, Logical Confusion, and Emotional Manipulation. As illustrated in Fig. 6, the model exhibits significant and consistent performance gains across all categories. Specifically, the mean accuracy rises from 0.82 to 0.90 for Fact Distortion, from 0.69 to 0.83 for Logical Confusion, and shows the most substantial improvement for Emotional Manipulation, increasing from an initial 0.59 to 0.82. More importantly, the performance disparity among these groups, measured by the inter-group gap, dramatically narrows from 0.23 to 0.08. This convergence indicates that InfoDirichlet Resample effectively prevents the model from overfitting to easier attack types. Instead, it progressively reallocates the learning focus toward the more challenging, weaker groups, ultimately yielding a more balanced and stable robustness against diverse and complex adversarial comment mechanisms.

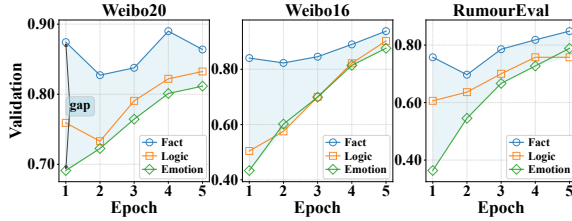


Figure 6: Validation accuracy across epochs on the three attack-type validation sets.

RQ3. Adaptive allocation progressively balances robustness across heterogeneous attack groups.

5.6 Ablation Study

To quantify the contribution of each component in CogniDir, we conduct an ablation study on Malicious Comments Generation (G), Vulnerability Score (VS), and Dirichlet Expectation Allocation (DEA). Based on Tab. 3, we show the model’s performance across the three datasets when removing a single module. Compared with the full model, removing G (CogniDir-G) yields the largest performance drop, with an average macro-

F1 decrease of 21.4% points. Replacing VS with random scores (CogniDir-VS) results in an average decrease of 16.3% points, while substituting DEA with a softmax-weighted scheme (CogniDir-DEA) also leads to a clear degradation, reducing macro-F1 by 11.1% points on average. These results indicate that the three components are essential and complementary for robust adversarial training, and removing any of them substantially weakens overall performance.

RQ3. Generation, vulnerability estimation, and adaptive allocation are complementary and all materially contribute to robustness.

Table 3: Evaluation results of ablating G, VS and DEA from CogniDir on both datasets.

Dataset	Method	mac F1	Acc.	F1-real	F1-fake
Weibo16	CogniDir-G	0.792	0.811	0.830	0.754
	CogniDir-VS	0.821	0.836	0.862	0.780
	CogniDir-DEA	0.855	0.872	0.895	0.815
	CogniDir	0.945	0.947	0.957	0.933
Weibo20	CogniDir-G	0.728	0.742	0.770	0.686
	CogniDir-VS	0.792	0.807	0.825	0.759
	CogniDir-DEA	0.834	0.848	0.872	0.796
	CogniDir	0.947	0.947	0.951	0.943
RumourEval-19	CogniDir-G	0.552	0.574	0.620	0.484
	CogniDir-VS	0.613	0.628	0.662	0.564
	CogniDir-DEA	0.692	0.707	0.748	0.636
	CogniDir	0.812	0.836	0.883	0.741

6 Conclusion

In this study, we design and evaluate three categories of malicious comments grounded in cognitive psychology: fact distortion, logical confusion, and emotional manipulation. Based on these mechanisms, we further propose CogniDir, which dynamically estimates group-wise vulnerability and adjusts training exposure toward more vulnerable attack types. Experimental results across three benchmark datasets demonstrate that CogniDir achieves strong robustness and detection performance under heterogeneous malicious-comment attacks. Further analyses show that adaptive allocation progressively reduces the vulnerability gap across attack groups, while the ablation study confirms the complementary contributions of malicious-comment generation, vulnerability estimation, and adaptive allocation. These findings indicate that explicitly modeling heterogeneous attack mechanisms is important for robust comment-aware fake news detection. Overall, CogniDir provides a practical framework for mechanism-aware and feedback-driven robust training in evolving adversarial social media environments.

Limitations

Despite the effectiveness of our proposed CogniDir, there are several limitations. First, our current evaluation primarily focuses on mechanism-controlled malicious comments synthesized by multiple LLMs. Although mixed-generator synthesis and human evaluation improve the diversity and quality of these comments, they cannot fully represent naturally occurring malicious comments in real-world social media. Second, we do not exhaustively evaluate attacks generated by unseen LLMs or unseen prompting strategies. Future work will extend CogniDir to naturally occurring and out-of-distribution comment attacks to further assess its robustness under more diverse adversarial settings.

Acknowledgements

This work was supported by the National Natural Science Foundation of China under Grants 62376265, 62502514, and 62372454.

References

- Christopher M. Bishop. 2006. *Pattern recognition and machine learning*. Springer.
- Canyu Chen and Kai Shu. 2024. Combating misinformation in the age of llms: Opportunities and challenges. *AI Magazine*, 45(3):354–368.
- Thomas M Cover and Joy A. Thomas. 1991. *Elements of information theory*. Wiley New York.
- Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and harnessing adversarial examples. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.
- Genevieve Gorrell, Elena Kochkina, Maria Liakata, Ahmet Aker, Arkaitz Zubiaga, Kalina Bontcheva, and Leon Derczynski. 2019. Semeval-2019 task 7: Rumoureal, determining rumour veracity and support for rumours. In *Proceedings of the 13th international workshop on semantic evaluation*, pages 845–854.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. *On calibration of modern neural networks*. Preprint, arXiv:1706.04599.
- Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. 2022. A survey on automated fact-checking. *Transactions of the association for computational linguistics*, 10:178–206.
- Beizhe Hu, Qiang Sheng, Juan Cao, Yang Li, and Danding Wang. 2025. Llm-generated fake news induces truth decay in news ecosystem: A case study on neural news recommendation. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 435–445.
- Beizhe Hu, Qiang Sheng, Juan Cao, Yuhui Shi, Yang Li, Danding Wang, and Peng Qi. 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, pages 22105–22113.
- Binyuan Hui, Jian Yang, Zeyu Cui, Jiaxi Yang, Dayiheng Liu, Lei Zhang, Tianyu Liu, Jiajun Zhang, Bowen Yu, Keming Lu, and 1 others. 2024. Qwen2.5-coder technical report. *arXiv preprint arXiv:2409.12186*.
- Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, and 1 others. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.
- Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, L lio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, and 7 others. 2024. *Mixtral of experts*. Preprint, arXiv:2401.04088.
- Annat Koren, Chandler Underwood, Edoardo Serra, and Francesca Spezzano. 2025. Evaluating the robustness of fake news detectors to adversarial attacks with real user comments. *International Journal of Data Science and Analytics*, 20(6):5689–5702.
- Thai Le, Suhang Wang, and Dongwon Lee. 2020. Malcom: Generating malicious comments to attack neural fake news detection models. In *2020 IEEE International Conference on Data Mining (ICDM)*, pages 282–291. IEEE.
- Zhiwei Liu, Boyang Liu, Paul Thompson, Kailai Yang, and Sophia Ananiadou. 2024. Conspemollm: Conspiracy theory detection using an emotion-based large language model. *arXiv preprint arXiv:2403.06765*.
- Jing Ma, Wei Gao, Prasenjit Mitra, Sejeong Kwon, Bernard J. Jansen, Kam-Fai Wong, and Meeyoung Cha. 2016. Detecting rumors from microblogs with recurrent neural networks. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI 2016, New York, NY, USA*,

- 9-15 July 2016, pages 3818–3824. IJCAI/AAAI Press.
- Milad Moradi and Matthias Samwald. 2021. Evaluating the robustness of neural language models to input perturbations. *arXiv preprint arXiv:2108.12237*.
- Qiong Nan, Qiang Sheng, Juan Cao, Beizhe Hu, Danding Wang, and Jintao Li. 2024. Let silence speak: Enhancing fake news detection with generated comments from large language models. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 1732–1742.
- Qiong Nan, Qiang Sheng, Juan Cao, Yongchun Zhu, Danding Wang, Guang Yang, and Jintao Li. 2025. Exploiting user comments for early detection of fake news prior to users’ commenting. *Frontiers of Computer Science*, 19(10):1910354.
- Richard E Nisbett, Kaiping Peng, Incheol Choi, and Ara Norenzayan. 2001. Culture and systems of thought: holistic versus analytic cognition. *Psychological review*, 108(2):291.
- Noel Nutsugah, Kobby Mensah, Raphael Odoom, and Amin Ayarnah. 2025. Social media, misinformation and fake news in the pandemic: the dominant gaps and future research avenues. *Online Information Review*, 49(2):335–352.
- Hame Park, Ayelet Arazi, Bharath Chandra Talluri, Marco Celotto, Stefano Panzeri, Alan A Stocker, and Tobias H Donner. 2025a. Confirmation bias through selective readout of information encoded in human parietal cortex. *Nature Communications*, 16(1):5391.
- Sungwon Park, Sungwon Han, Xing Xie, Jae-Gil Lee, and Meeyoung Cha. 2025b. Adversarial style augmentation via large language model for robust fake news detection. In *Proceedings of the ACM on Web Conference 2025*, pages 4024–4033.
- Gordon Pennycook and David G Rand. 2021. The psychology of fake news. *Trends in cognitive sciences*, 25(5):388–402.
- Piotr Przybyła, Euan McGill, and Horacio Saggion. 2024. Attacking misinformation detection using adversarial examples generated by language models. *ArXiv preprint*, abs/2410.20940.
- Dongning Rao, Xin Miao, Zhihua Jiang, and Ran Li. 2021. Stanker: Stacking network based on level-grained attention-masked bert for rumor detection on social media. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pages 3347–3363.
- Kai Shu, Limeng Cui, Suhang Wang, Dongwon Lee, and Huan Liu. 2019. defend: Explainable fake news detection. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 395–405.
- Zixun Sun, Mingye Xu, Guanming Liang, and Qi Liu. 2025. Unifying adversarial multi-deconfounded learning paradigm for fake news detection. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1*, pages 2572–2583.
- Henri Tajfel, Michael G Billig, Robert P Bundy, and Claude Flament. 1971. Social categorization and intergroup behaviour. *European journal of social psychology*, 1(2):149–178.
- Zhenhua Tan and Tao Zhang. 2026. Emotion-semantic interaction network for fake news detection: Perspectives on question and non-question comment semantics. *Information Processing & Management*, 63(2):104391.
- Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, and 1 others. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.
- Zhao Tong, Chunlin Gong, Yiping Zhang, Haichao Shi, Qiang Liu, Xingcheng Xu, Shu Wu, and Xiao-Yu Zhang. 2026a. Cot is not the chain of truth: An empirical internal analysis of reasoning llms for fake news generation. *arXiv preprint arXiv:2602.04856*.
- Zhao Tong, Yimeng Gu, Huidong Liu, Qiang Liu, Shu Wu, Haichao Shi, and Xiao-Yu Zhang. 2025. Generate first, then sample: Enhancing fake news detection with llm-augmented reinforced sampling. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 24276–24290.
- Zhao Tong, Pengfei Yang, Yimeng Gu, Haichao Shi, Qiang Liu, Xingcheng Xu, Shu Wu, and Xiao-Yu Zhang. 2026b. Rlshield: Dynamic jailbreak detection for llms via reinforced adaptive learning. In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 23615–23630.
- Amos Tversky and Daniel Kahneman. 1978. [2 - judgment under uncertainty: Heuristics and biases: Biases in judgments reveal some heuristics of thinking under uncertainty](#). In PETER DIAMOND and MICHAEL ROTHCHILD, editors, *Uncertainty in Economics*, pages 17–34. Academic Press.
- Amos Tversky and Daniel Kahneman. 1981. The framing of decisions and the psychology of choice. *science*, 211(4481):453–458.
- Herun Wan, Shangbin Feng, Zhaoxuan Tan, Heng Wang, Yulia Tsvetkov, and Minnan Luo. 2024. Dell: Generating reactions and explanations for llm-based misinformation detection. *ArXiv preprint*, abs/2402.10426.
- Bo Wang, Jing Ma, Hongzhan Lin, Zhiwei Yang, Ruichao Yang, Yuan Tian, and Yi Chang. 2024a. Explainable fake news detection with large language

- model via defense among competing wisdom. In *Proceedings of the ACM Web Conference 2024*, pages 2452–2463.
- Jinguang Wang, Shengsheng Qian, Jun Hu, and Richang Hong. 2023. Comment-context dual collaborative masked transformer network for fake news detection. *IEEE Transactions on Multimedia*, 26:5170–5180.
- Lanjuan Wang, Zehao Wang, Le Wu, and An-An Liu. 2024b. Bots shield fake news: adversarial attack on user engagement based fake news detection. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 2369–2378.
- Jiaying Wu, Jiafeng Guo, and Bryan Hooi. 2024. Fake news in sheep’s clothing: Robust fake news detection against llm-empowered style attacks. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD 2024, Barcelona, Spain, August 25-29, 2024*, pages 3367–3378. ACM.
- Yuta Yanagi, Ryohei Orihara, Yuichi Sei, Yasuyuki Tahara, and Akihiko Ohsuga. 2020. Fake news detection with generated comments for news articles. In *2020 IEEE 24th International Conference on Intelligent Engineering Systems (INES)*, pages 85–90. IEEE.
- An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, and 1 others. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.
- Chang Yang, Peng Zhang, Wenbo Qiao, Hui Gao, and Jiaming Zhao. 2023a. Rumor detection on social media with crowd intelligence and ChatGPT-assisted networks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5705–5717. Association for Computational Linguistics.
- Sin-Han Yang, Chung-Chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2023b. Entity-aware dual co-attention network for fake news detection. *arXiv preprint arXiv:2302.03475*.
- Min-Hsuan Yeh, Ruyuan Wan, and Ting-Hao Huang. 2024. Cocolofa: A dataset of news comments with common logical fallacies written by llm-assisted crowds. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 660–677.
- Zepu Yi, Chenxu Tang, and Songfeng Lu. 2025. User comment-guided cross-modal attention for interpretable multimodal fake news detection. *Applied Sciences*, 15(14):7904.
- Peican Zhu, Zechen Pan, Yang Liu, Jiwei Tian, Keke Tang, and Zhen Wang. 2024. [A general black-box adversarial attack on graph-based fake news detectors](#). *Preprint*, arXiv:2404.15744.

Appendix Overview

The appendix is organized as follows:

- **Appendix A** Threat model and direction-specific attack analysis.
- **Appendix B** Original dataset annotation schemes.
- **Appendix C** Generated comment construction and quality analysis.
- **Appendix D** Generator selection and bias analysis.
- **Appendix E** Human evaluation.
- **Appendix F** Real-world mechanism correspondence.
- **Appendix G** Derivation for vulnerability score.
- **Appendix H** Theoretical view of adaptive allocation.
- **Appendix I** Baselines and evaluation protocols.
- **Appendix J** Statistical analysis across multiple runs.

A Threat Model and Direction-Specific Attack Analysis

We consider an untargeted prediction-manipulation attack against comment-aware fake news detectors. The attacker keeps the news content unchanged and injects malicious comments into the associated comment set. An attack succeeds whenever an originally correct prediction becomes incorrect, without specifying a target label in advance.

Accordingly, the overall attack success rate includes both real-to-fake and fake-to-real prediction flips. For the two directions, we define

$$\text{ASR}_{R \rightarrow F} = \frac{\sum_i \mathbb{I}[y_i = 0 \wedge \hat{y}_i^O = 0 \wedge \hat{y}_i^A = 1]}{\sum_i \mathbb{I}[y_i = 0 \wedge \hat{y}_i^O = 0]}, \quad (11)$$

and

$$\text{ASR}_{F \rightarrow R} = \frac{\sum_i \mathbb{I}[y_i = 1 \wedge \hat{y}_i^O = 1 \wedge \hat{y}_i^A = 0]}{\sum_i \mathbb{I}[y_i = 1 \wedge \hat{y}_i^O = 1]}. \quad (12)$$

Table 4: Direction-specific attack success rates(ASR) on GenFEND.

Dataset	F1 under Attack	$R \rightarrow F$	$F \rightarrow R$
Weibo16	0.578	36.4%	47.9%
Weibo20	0.640	39.0%	31.6%
RumourEval-19	0.456	54.5%	50.0%

Table 5: Annotation schemes provided by the original datasets.

Dataset	Veracity	Comment Label	Mechanism Label
Weibo16	Yes	None	None
Weibo20	Yes	None	None
RumourEval-19	Yes	Stance	None

Using GenFEND as a representative baseline, Tab. 4 reports the attack effectiveness in both directions.

The substantial attack success rates in both directions indicate that the observed degradation reflects a general vulnerability to malicious comment manipulation instead of an attack specialized toward a single target label.

B Original Dataset Annotation Schemes

The original datasets primarily provide task-level annotations for fake news detection. Weibo16 and Weibo20 provide news-level veracity labels, while RumourEval-19 additionally provides reply-level stance labels (as shown in Tab.5). None of the three datasets contains annotations for the cognitive manipulation mechanisms studied in this work.

Consequently, the original annotations cannot directly characterize the cognitive manipulation embodied in malicious comments or distinguish the effects of different attack mechanisms.

C Generated Comment Construction and Quality

C.1 Prompt Construction Protocol

The attack-generation prompt is derived from the task definition rather than selected according to downstream detection performance. It follows an understand-first, then-generate procedure. The first stage identifies salient facts in the news and determines a misleading direction consistent with the specified attack mechanism. The second stage generates a comment conditioned on the news context, attack mechanism, and natural user tone.

Table 6: Lexical diversity of generated malicious comments.

Dataset	Distinct-1	Distinct-2	Distinct-3
Weibo16	0.29	0.58	0.76
Weibo20	0.31	0.60	0.78
RumourEval-19	0.27	0.55	0.73

The prompt template was fixed before large-scale generation and applied consistently across datasets and generation models. No prompt variants were selected based on downstream detection results.

C.2 Lexical Diversity via Distinct- n

We compute Distinct- n for $n \in \{1, 2, 3\}$ on the generated comments to examine lexical diversity.

C.3 Standardized OOD Distance

We further compare generated comments with original comments in representation space. Given a fixed multilingual sentence encoder ϕ , we compute the centroid μ_{real} of the original comments. The cosine distance of a comment c from this centroid is

$$d(c) = 1 - \cos(\phi(c), \mu_{\text{real}}). \quad (13)$$

Using the mean μ_d and standard deviation σ_d of the distances of original comments, we define the standardized distance as

$$d_{\text{std}}(c) = \frac{d(c) - \mu_d}{\sigma_d}. \quad (14)$$

The mean standardized distances are 0.25, 0.32, and 0.35 for Fact Distortion, Logical Confusion, and Emotional Manipulation, respectively.

C.4 Case Study: Generated Adversarial Comments

To illustrate the distinctions among the three attack mechanisms, we present representative generated comments conditioned on the same news example.

News Sample (Label: Real)

News Content: “Local health bureau: No new respiratory illness cases this week. All patients recovered. Continue wearing masks in crowded areas.”

Table 7: Macro-F1 under different attack-generator settings.

Dataset	Original	Qwen-32B Robust	Mixed Robust
Weibo16	0.955	0.948	0.945
Weibo20	0.936	0.951	0.947
RumourEval-19	0.822	0.924	0.812

Fact Distortion

Generated Comment: “I checked the hospital WeChat—they actually reported 47 new cases last week but only counted recovered ones to keep the number low!”

Why misleading: The comment fabricates factual details and numerical evidence, creating false certainty and undermining trust in the reported statistics.

Logical Confusion

Generated Comment: “If there are really zero new cases, why are they still telling us to wear masks everywhere? Either the treatment doesn’t work or the data is fake.”

Why misleading: The comment introduces flawed causal reasoning and a false dichotomy that creates doubt while appearing superficially coherent.

Emotional Manipulation

Generated Comment: “My friend in the hospital says they’re hiding deaths to avoid panic before the holiday... Don’t believe the official numbers, protect your family!”

Why misleading: The comment uses fear, conspiracy framing, and personal anecdotes to encourage authenticity judgments driven by emotion instead of verifiable evidence.

D Generator Selection and Bias Analysis

The main experiments combine malicious comments generated by Mistral-7B, Gemma-2B, and Qwen-32B. This mixed-generator design is intended to reduce the possibility that the detector learns generator-specific vocabulary, sentence patterns, or semantic styles.

To examine this issue, we compare the mixed-generator setting with a single-generator setting in which Qwen-32B is used for both robust-training and test attack comments.

The single-generator setting produces an unusually large result on RumourEval-19, where robust performance reaches 0.924 compared with an original performance of 0.822. Because the same generator is used for training and testing in this setting, this improvement may reflect generator-specific shortcuts in addition to genuine robustness recovery. The mixed-generator setting does not exhibit the same anomalous increase. We therefore use mixed generation in the main experiments to reduce dependence on the stylistic characteristics of a single generator.

E Human Evaluation

We conduct a stratified human evaluation of the generated malicious comments. Samples are stratified by dataset (WEIBO16, WEIBO20, and RUMOUREVAL-19), veracity label (real or fake), and attack mechanism (Fact Distortion, Logical Confusion, and Emotional Manipulation), yielding 18 strata. We uniformly sample 30 comments from each stratum, resulting in $N = 540$ evaluated comments.

Each comment is independently assessed by two graduate-level annotators according to four criteria:

- **Relevance:** The comment is on-topic and refers to the entities or events described in the news.
- **Naturalness:** The comment resembles plausible user-generated content in tone and expression.
- **Attack-Type Fidelity:** The comment is consistent with the intended cognitive attack mechanism.
- **Misleading Potential:** The comment can plausibly interfere with authenticity judgment.

A sample passes the quality check when it satisfies relevance, naturalness, and attack-type fidelity. Disagreements are resolved through discussion or additional adjudication. Inter-annotator agreement on acceptance is substantial, with Cohen’s $\kappa = 0.72$.

The high acceptance rates across all three mechanisms support the linguistic quality and mechanism fidelity of the generated comments. Rejected samples are regenerated once.

Table 8: Human-evaluation acceptance rates.

Attack Mechanism	Acceptance Rate
Fact Distortion	92.1%
Logical Confusion	90.3%
Emotional Manipulation	89.4%
Overall	90.6%

F Real-World Mechanism Correspondence

To examine whether the three mechanisms correspond to patterns observed in real-world misleading user content, we qualitatively compare generated attack comments with representative real-world cases.

These examples illustrate qualitative correspondence between the mechanisms instantiated by our generated comments and misleading patterns observed in real-world user content. They do not constitute a representative benchmark of naturally occurring attacks, which remains an important direction for future evaluation.

G Derivation for Vulnerability Score

For attack group j , let

$$\hat{p}_i = p_\theta(y_i \mid x_i, \mathcal{C}_i) \quad (15)$$

denote the probability assigned to the true label for validation sample i . The average cross-entropy loss is

$$m_{j,\text{loss}} = -\frac{1}{n} \sum_{i=1}^n \log \hat{p}_i. \quad (16)$$

The corresponding confidence proxy is

$$q_j = \exp(-m_{j,\text{loss}}) = \left(\prod_{i=1}^n \hat{p}_i \right)^{1/n}, \quad (17)$$

which is the geometric mean of the true-label probabilities.

Let

$$a_j = m_{j,\text{acc}} \quad (18)$$

denote empirical accuracy for attack group j . We define the group-wise vulnerability score as

$$s_j = -a_j \log q_j - (1 - a_j) \log(1 - q_j). \quad (19)$$

Table 9: Qualitative correspondence between real-world misleading content and generated attack comments.

Case	Real-World Misleading Content	Generated Attack Comment	Mechanism Correspondence
“Xu Jing” family-search post during the Ya’an Lushan earthquake	A reposted help message used a specific name, hospital, location, and contact information to create credibility and encourage further dissemination.	“This urgent request for help appears highly credible and provides specific details, including a name, location, and phone number. I hope Xu Jing receives the message soon. Please help spread it!”	Fact Distortion. Both rely on specific but unverified details to create false credibility and encourage dissemination.
Rapid increase in signatures on a South Korean Blue House national petition	A rapid increase in comments and petition recommendations was interpreted as evidence that unspecified external forces were manipulating the process.	“The surging number of petition signatures raises doubts about its authenticity. Receiving more than 17,000 signatures overnight is usually associated with data manipulation.”	Logical Confusion. Both infer deliberate manipulation directly from a short-term increase in quantity without intermediate evidence.
“Crisis actor” narrative surrounding the Boston Marathon bombing	Inconsistent footage and eyewitness accounts were framed as evidence that the event had been staged.	“The victims in the incident were actually U.S. soldiers who lost limbs in Afghanistan, and the scene appears to have been carefully staged, which deserves further thought.”	Emotional Manipulation. Both use conspiratorial narratives to induce shock and distrust and to shift judgment away from verifiable evidence.

This score can be decomposed as

$$s_j = H(\text{Bern}(a_j)) + D_{\text{KL}}(\text{Bern}(a_j) \parallel \text{Bern}(q_j)). \quad (20)$$

The entropy term characterizes uncertainty in empirical correctness, while the KL term captures the mismatch between empirical correctness and the model-implied true-label confidence. The score therefore couples accuracy and confidence in a fixed form and does not introduce a manually tuned combination weight.

H Theoretical View of Adaptive Allocation

H.1 Dirichlet-Mean Parameterization

For each attack group $j \in \{F, L, E\}$, we map the vulnerability score to a positive parameter

$$\alpha_j = s_j + 1. \quad (21)$$

The group allocation is defined using the deterministic Dirichlet mean

$$\bar{p}_j = \frac{\alpha_j}{\sum_k \alpha_k} = \frac{s_j + 1}{\sum_k (s_k + 1)}. \quad (22)$$

This parameterization maps group-wise vulnerability scores to a valid probability vector on the simplex.

H.2 Entropy-Regularized Allocation View

Let

$$u_j = \log \alpha_j \quad (23)$$

and let

$$\Delta = \left\{ p \mid p_j \geq 0, \sum_j p_j = 1 \right\} \quad (24)$$

denote the probability simplex. The allocation can equivalently be expressed as the solution of

$$p^* = \arg \max_{p \in \Delta} \left[\sum_j p_j u_j + H(p) \right], \quad (25)$$

where

$$H(p) = - \sum_j p_j \log p_j \quad (26)$$

is Shannon entropy. The unique closed-form solution is

$$p_j^* = \frac{\exp(u_j)}{\sum_k \exp(u_k)} = \frac{\alpha_j}{\sum_k \alpha_k} = \bar{p}_j. \quad (27)$$

Thus, the deterministic allocation used in CogniDir can be interpreted as an entropy-regularized group-allocation rule with a unique solution.

H.3 Allocation Properties

Because $\alpha_j = s_j + 1 > 0$, every group receives positive probability,

$$\bar{p}_j > 0. \quad (28)$$

The probabilities satisfy

$$\sum_j \bar{p}_j = 1. \quad (29)$$

Let $S = \sum_k \alpha_k$. The response of a group’s allocation to its own vulnerability score is

$$\frac{\partial \bar{p}_j}{\partial s_j} = \frac{S - \alpha_j}{S^2} > 0, \quad (30)$$

showing that increasing the vulnerability score of group j , while holding the other group scores fixed, increases its training allocation.

These properties guarantee simplex feasibility, positive allocation, and monotonic response to group vulnerability. They do not imply a globally optimal training-data mixture for the downstream learning problem.

H.4 Relation to Dirichlet Modeling

The Dirichlet distribution is used as a convenient parameterization of the group probability simplex. CogniDir does not define a Bayesian prior over model parameters, perform prior-posterior updating, or sample a posterior distribution. Training uses the deterministic Dirichlet mean rather than random draws from $\text{Dir}(\alpha)$.

Moreover, the implemented allocation does not introduce an independently tuned Dirichlet concentration hyperparameter. The group parameters are determined directly from the vulnerability scores at each training round.

Given a total comment budget M , the expected group allocation is converted into integer sample counts using

$$M_j = \lfloor \bar{p}_j M \rfloor, \quad (31)$$

with any remaining samples assigned according to the largest fractional remainders.

I Baselines and Evaluation Protocols

We compare against two categories of baselines: (1) LLM-only zero-shot detectors and (2) deep-learning comment-aware fake-news detection models. All baselines receive the news article and its associated comment set as input and output a binary prediction (Fake or Real).

I.1 LLM-Only Baselines

We evaluate four instruction-following large language models in a zero-shot setting without task-specific finetuning on the target datasets:

- **Gemma-2-2B** (Team et al., 2024): A lightweight open-weight language model designed for efficient language understanding and generation.
- **Mistral-7B** (Jiang et al., 2023): An efficient 7B-parameter language model using grouped-query attention and sliding-window attention.
- **Llama-3-8B** (Grattafiori et al., 2024): An instruction-following language model with

Table 10: Macro-F1 of LLM-only baselines under temperature 0.2 and greedy decoding ($T = 0$).

Dataset	Model	T	O	A	R
Weibo16	Gemma-2-2B	0.2	0.322	0.291	0.274
		0	0.514	0.292	0.309
	Mistral-7B	0.2	0.686	0.334	0.495
		0	0.668	0.372	0.401
	Llama-3-8B	0.2	0.598	0.324	0.408
		0	0.483	0.356	0.384
	Qwen2.5-32B	0.2	0.831	0.282	0.716
		0	0.805	0.495	0.529
Weibo20	Gemma-2-2B	0.2	0.345	0.292	0.326
		0	0.523	0.296	0.321
	Mistral-7B	0.2	0.654	0.315	0.375
		0	0.583	0.387	0.420
	Llama-3-8B	0.2	0.582	0.363	0.408
		0	0.423	0.337	0.357
	Qwen2.5-32B	0.2	0.823	0.367	0.541
		0	0.797	0.485	0.515
RumourEval-19	Gemma-2-2B	0.2	0.294	0.238	0.275
		0	0.441	0.238	0.258
	Mistral-7B	0.2	0.468	0.184	0.252
		0	0.549	0.310	0.339
	Llama-3-8B	0.2	0.377	0.219	0.276
		0	0.469	0.199	0.226
	Qwen2.5-32B	0.2	0.553	0.343	0.412
		0	0.523	0.352	0.377

strong general language and reasoning capabilities.

- **Qwen2.5-32B** (Yang et al., 2024): A multilingual instruction-following language model with strong reasoning and structured-generation capabilities.

I.2 Inference Reproducibility

The main experiments use a temperature of 0.2 for all LLM-only baselines. To examine sensitivity to decoding randomness, we additionally evaluate all four models using greedy decoding ($T = 0$) while keeping the prompts, input samples, and evaluation protocol unchanged.

As shown in Tab. 10, the overall conclusions remain stable across the two decoding settings. The mean Macro-F1 differences are 0.004, 0.005, and 0.033 on Weibo16, Weibo20, and RumourEval-19, respectively, with an overall mean difference of 0.014.

I.3 Deep-Learning Baselines

We include five representative comment-aware deep-learning baselines:

- **dDEFEND** (Shu et al., 2019): An explainable model using sentence-comment co-attention to jointly model news and comment evidence.

- **Dual-CAN** (Yang et al., 2023b): An entity-aware dual co-attention network that models news content, user replies, and complementary evidence.
- **GenFEND** (Nan et al., 2024): A generative feedback-enhanced detector that leverages LLM-generated comments to enrich detection evidence.
- **L-Defense** (Wang et al., 2024a): An LLM-based explainable detector that extracts competing evidence for authenticity prediction.
- **ARG** (Hu et al., 2024): A representative comment-aware fake-news detection baseline included in our evaluation.

I.4 Evaluation Protocol

Throughout the evaluation, O denotes performance on the original comment set. A denotes performance on attacked inputs using a model without robust training. R denotes performance after robust training on the same attacked test inputs used for A. Therefore, A and R are evaluated under matched adversarial inputs.

I.5 Metrics

- **Accuracy (Acc):** Overall correctness on the test set.

$$\text{Acc} = \frac{1}{N} \sum_{i=1}^N \mathbb{I}[\hat{y}_i = y_i]. \quad (32)$$

- **Macro-F1:** Unweighted average of the class-wise F1 scores.

$$\text{Macro-F1} = \frac{1}{2} (F1_{\text{Real}} + F1_{\text{Fake}}). \quad (33)$$

- **Attack Success Rate (ASR):** Fraction of originally correctly classified instances that become misclassified after malicious comments are introduced.

$$\text{ASR} = \frac{\sum_{i=1}^N \mathbb{I}[\hat{y}_i^O = y_i \wedge \hat{y}_i^A \neq y_i]}{\sum_{i=1}^N \mathbb{I}[\hat{y}_i^O = y_i]}. \quad (34)$$

Table 11: Mean $\pm$ sample standard deviation over three runs.

Dataset	Model	O	A	R
Weibo16	GenFEND	0.921 $\pm$ 0.027	0.577 $\pm$ 0.025	0.862 $\pm$ 0.019
	CogniDir	0.954 $\pm$ 0.010	0.789 $\pm$ 0.020	0.943 $\pm$ 0.011
Weibo20	GenFEND	0.889 $\pm$ 0.020	0.642 $\pm$ 0.013	0.761 $\pm$ 0.021
	CogniDir	0.934 $\pm$ 0.012	0.729 $\pm$ 0.020	0.947 $\pm$ 0.012
RumourEval-19	GenFEND	0.662 $\pm$ 0.017	0.457 $\pm$ 0.024	0.603 $\pm$ 0.020
	CogniDir	0.823 $\pm$ 0.015	0.549 $\pm$ 0.019	0.810 $\pm$ 0.015

Table 12: Two-sided paired t -tests between CogniDir and GenFEND.

Dataset	Setting	Δ	p -value
Weibo16	O	+0.033	0.089
	A	+0.212	< 0.001
	R	+0.080	0.042
Weibo20	O	+0.045	0.011
	A	+0.087	0.003
	R	+0.185	0.010
RumourEval-19	O	+0.161	0.012
	A	+0.092	0.063
	R	+0.207	0.009

J Statistical Analysis

J.1 Three-Run Stability

We conduct two additional independent runs for CogniDir and GenFEND under matched experimental settings. Together with the original run, Tab. 11 reports the mean and sample standard deviation over three runs.

CogniDir achieves a higher three-run mean in all nine dataset-setting combinations. The relatively small standard deviations also indicate that the main performance trends are stable across runs.

J.2 Paired Significance Tests

We further conduct two-sided paired t -tests over the three matched runs.

All nine mean differences are positive. Seven of the nine comparisons are statistically significant at the uncorrected $p < 0.05$ level. The improvements on Weibo16-O and RumourEval-19-A remain positive but do not reach this significance threshold.