Fake News Court: A Multi-Agent Adversarial Framework for Robust Detection of LLM-Generated Fake News

Anonymous ACL submission

Abstract

Large language models (LLMs) have achieved remarkable success across many domains, yet their generative capabilities have been misused to produce highly realistic fake news that threatens social stability. However, the main existing detection methods, including small language model (SLM)-based approaches and LLM-based detectors, exhibit substantial performance degradation on LLM-generated content, reflecting a lack of robustness. SLM-based detectors rely heavily on fixed data distributions and shallow textual cues, while LLM-based detectors depend on prompt-based inference, making them sensitive to prompting and vulnerable to LLM hallucinations. To address this challenge, we propose Fake News Court (FNC), a robust multi-agent adversarial framework that integrates the complementary strengths of LLM and SLM for detecting LLM-generated fake news. FNC adopts an adversarial learning paradigm in which a quality-controlled generator produces diverse and challenging fake news, while a hybrid detector combines multi-dimensional LLM-based agent reasoning with an SLM-based classifier to ensure stable decisions. Extensive experiments show that FNC improves detection accuracy by an average of 6% on LLM-generated fake news compared to the state-of-the-art method, confirming its robustness.

1 Introduction

In recent years, large language models (LLMs) have delivered transformative progress across many domains due to their rich world knowledge and strong language understanding and reasoning (Maddigan and Susnjak, 2023). However, their rapidly improved text generation has also lowered the barrier for producing highly realistic fake news, which is harder to detect than human-written misinformation and poses a growing threat to social stability (Hu et al., 2025). In this context, existing fake news detection methods struggle to reliably identify LLM-generated fake news, as illustrated in Figure 1(a).

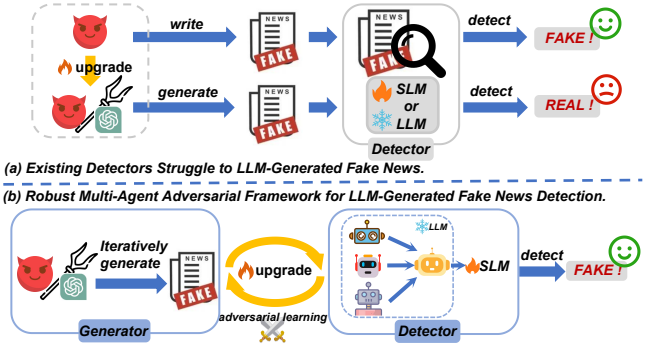

Figure 1: Comparison between existing fake news detection methods and FNC.

Specifically, existing methods mainly include approaches based on SLM and relying on LLM. The SLM-based methods tend to rely on historical data distributions and surface-level textual patterns, which fail to generalize to the fluent, diverse, and adaptive content produced by LLMs (Ali et al., 2025). Meanwhile, LLM-based detectors commonly adopt prompt-based self-evaluation and improve performance through prompt optimization, where similar or identical models are used for both generation and judgment (Wang et al., 2025). However, constrained by limited supervision samples and the inherent boundaries of LLM knowledge, such detectors tend to be brittle under distribution shifts and exhibit weak detection capability on out-of-distribution fake news. These limitations underscore the need for a detection framework capable of robustly handling LLM-generated fake news.

To address this challenge, we propose Fake News Court (FNC), a robust multi-agent adversarial framework for LLM-generated fake news detection. FNC couples a generator (FNC-G) and a detector (FNC-D) and trains them through adversarial interaction as illustrated in Figure 1(b). FNC-G adopts an adversarial data generation paradigm to

synthesize diverse and complex out-of-distribution samples, reducing the detector's reliance on in-distribution data and the LLM's intrinsic knowledge. Meanwhile, for the detector, we leverage the reasoning capability of LLMs and integrate the more robust learning ability of SLMs to improve detection robustness.

In FNC-G, we generate diverse adversarial samples via two complementary agents. A label-preserving agent rewrites the input while keeping its veracity label unchanged, with prompts controlling the extent of surface modification. A label-flipping agent starts from real news and introduces controlled modifications to key factual statements or semantic relations, ranging from mild perturbations to more substantial alterations while maintaining topical coherence and contextual relevance. To avoid hallucinated or off-topic outputs during training, an alignment gate filters candidates based on contextual, semantic, and factual consistency with the source, while more open-ended LLM-generated fake news is further considered during robustness evaluation.

In FNC-D, we adopt a hybrid reasoning and judgment design. LLM-based reviewer agents are organized into dimension-specific groups for factual consistency, semantic plausibility, and linguistic style. Within each group, heterogeneous reasoning strategies and structured deliberation reduce bias from homogeneous LLM reasoning. A secretary agent then consolidates the multi-dimensional analyses into a standardized rationale, which is jointly considered with the original news by a fine-tuned SLM-based classifier to produce the final verdict.

Our contributions are summarized as follows:

- We propose a robust multi-agent adversarial framework that integrates LLM-based generation and reasoning with SLM-based judgment for detecting LLM-generated fake news.
- We introduce a quality-controlled generator with alignment gate-based filtering, together with a hybrid detector that separates multi-dimensional reasoning by LLM-based agents from SLM-based decision making for stable and robust fake news detection.
- Extensive experiments demonstrate that FNC achieves an average improvement of 6% in detection accuracy over the representative state-of-the-art baseline on LLM-generated fake news datasets.

2 Related Work

Early fake news detection evolved from traditional machine learning methods to deep neural architectures and pretrained SLMs such as BERT, which have become the dominant paradigm in recent years (Shu et al., 2017; Choraś et al., 2021; Khalil et al., 2021; Rana et al., 2022; Ni et al., 2021).

However, rapid progress in LLM-based text generation has made LLM-generated fake news increasingly difficult to distinguish from genuine content (Sallami et al., 2024). Existing detectors, including SLM-based approaches (Lan et al., 2019; Raffel et al., 2020) and prompt-driven LLM-based methods (Bhattacharjee and Liu, 2024; Wang et al., 2024b), often generalize poorly to LLM-generated misinformation due to distribution shifts, hallucination, and prompt sensitivity. Recent studies further introduce multi-agent discussion to improve robustness (Tian et al., 2025; Liu et al., 2025), but they still rely on simple aggregation of LLM outputs and provide limited control over adversarial sample quality. These limitations further highlight the urgent need for a robust LLM-generated fake news detection framework.

3 Methodology

3.1 Problem Formulation

We study the problem of robust detection of LLM-generated fake news, inspired by an adversarial training paradigm (Shafahi et al., 2019). Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$ denote a dataset of N human-written news articles, where x_i is a news text and $y_i \in \{\text{Real}, \text{Fake}\}$ is its ground-truth veracity label.

Given a sample (x, y), we consider a constrained generation process that produces an adversarial variant $\tilde{x} = \mathrm{G}(x \mid c)$ from x, where G denotes a generator and c represents constraints that enforce semantic validity, contextual coherence and factual consistency, ensuring that generated samples are suitable for robust detector training. The generated sample is associated with a label $\tilde{y}$, which may either preserve the original label ($\tilde{y} = y$) or correspond to a controlled label-flipped variant ($\tilde{y} \neq y$).

We further define a detector D_Φ parameterized by Φ, which takes an input news article $\tilde{x}$ and outputs a prediction $\hat{y} = \mathrm{D}_\Phi(\tilde{x})$.

3.2 Overall Framework

As shown in Figure 2, FNC consists of a generator (FNC-G) and a detector (FNC-D). Under adversarial training, detector is optimized to achieve

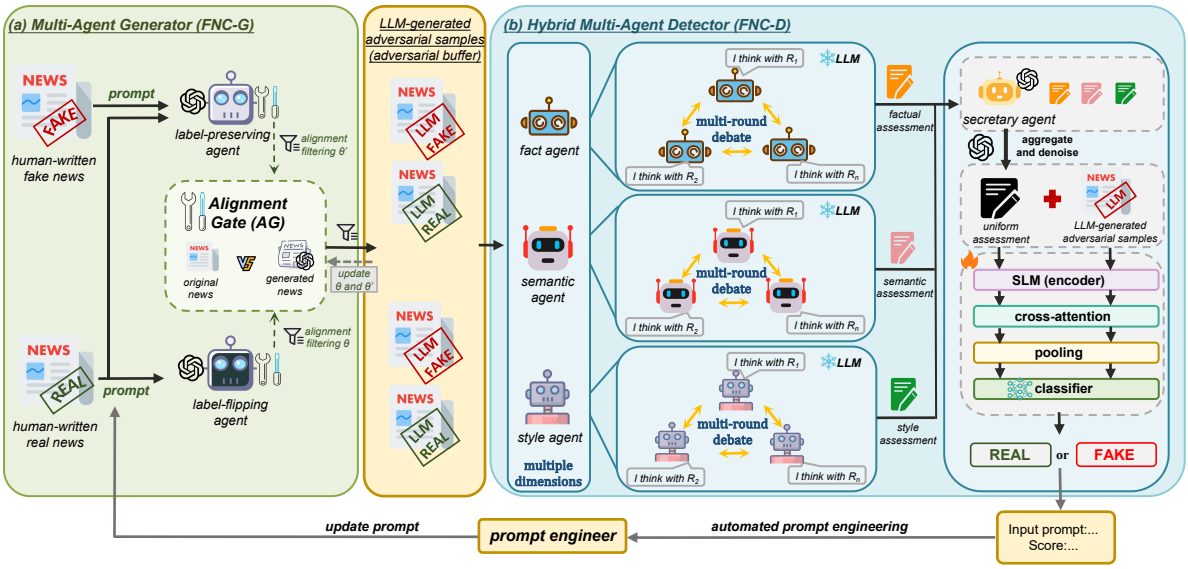

Figure 2: Overview of the proposed Fake News Court (FNC) framework.

accurate and stable predictions on adversarial samples, while the generation strategy is progressively refined to produce increasingly challenging yet valid adversarial news. Through this adversarial co-evolution, the detector is trained to be robust against evolving and diverse LLM-generated fake news distributions rather than fixed generation patterns.

FNC-G comprises a label-preserving agent and a label-flipping agent conditioned on human-labeled samples (x, y) to generate controllable adversarial samples, with an alignment gate filtering low-quality outputs and a prompt optimizer refining generation strategies. **FNC-D** adopts a hybrid design that combines multi-dimensional LLM-based agent deliberation with an SLM-based classifier to produce stable final predictions.

The detailed designs of FNC-G, FNC-D, and the adversarial training strategy are presented in the following sections.

3.3 Multi-Agent Generator (FNC-G)

FNC-G is powered by LLMs to generate diverse and high-quality adversarial samples for robust detector training. Following the problem formulation, we design two generator agents, namely a label-preserving agent and a label-flipping agent, whose detailed designs are described below.

Label-preserving Agent $\mathcal{A}_{\text{LP}}$. Existing studies commonly characterize LLM-generated fake news through rewriting and perturbation-based transformations rather than a unified generation taxonomy (Wang et al., 2024a; Chen and Shu, 2024; Zhou et al., 2023; Spitale et al., 2023). The label-preserving agent aims to generate alternative news texts that preserve the original semantics and factual content while modifying surface-level expressions, such that the veracity label remains unchanged. Guided by label-preserving prompts Π_{LP}, $\mathcal{A}_{\text{LP}}$ produces rewritten texts $\tilde{x}_{\text{LP}} = \mathcal{A}_{\text{LP}}(x \mid \Pi_{\text{LP}})$ with $\tilde{y} = y$. To capture different modification granularities, we employ prompts of two levels: light paraphrasing ($\mathbf{L}_0$) prompts $\pi^a_{\text{LP}} \in \Pi_{\text{LP}}$ introduce minor lexical or syntactic variations with minimal stylistic change, while stronger rewriting ($\mathbf{L}_1$) prompts $\pi^b_{\text{LP}} \in \Pi_{\text{LP}}$ allow more flexible re-expression of the content and style, provided that the core semantics and factual statements are preserved.

Label-flipping Agent $\mathcal{A}_{\text{LF}}$. The label-flipping agent aims to generate adversarial fake news by transforming human-written real news into label-inverted variants. Guided by label-flipping prompts Π_{LF}, $\mathcal{A}_{\text{LF}}$ generates perturbed texts $\tilde{x}_{\text{LF}} = \mathcal{A}_{\text{LF}}(x \mid \Pi_{\text{LF}})$ with $\tilde{y} \neq y$. To support different degrees of factual manipulation, we employ prompts with increasing levels: mild perturbation ($\mathbf{L}_2$) prompts $\pi^a_{\text{LF}} \in \Pi_{\text{LF}}$ introduce limited changes to key factual statements or semantic relations while preserving the overall discourse context, whereas stronger ($\mathbf{L}_3$) prompts $\pi^b_{\text{LF}} \in \Pi_{\text{LF}}$ allow more substantial factual changes or regeneration aligned with the original theme, while preserving topical coherence and narrative plausibility.

In practice, LLM safety mechanisms may restrict direct generation of adversarial content (Zhao et al.,

Agent Type	Representative Prompt Description
Label-preserving	($\mathbf{L}_0$) Persona: News editor. Paraphrase the article {news} by rephrasing sentences and expressions while preserving the original meaning, factual content, and journalistic style. ($\mathbf{L}_1$) Persona: Senior journalist. Rewrite the article {news} in a professional news tone, reorganizing expressions and sentence structures as needed, while strictly preserving the original facts.
Label-flipping	($\mathbf{L}_2$) Persona: Investigative journalist. Modify a small number of key factual statements in {news} to subtly alter its claims, while keeping the overall topic and writing style unchanged. ($\mathbf{L}_3$) Persona: Independent analyst. Based on the topic of {news}, rewrite the article into a misleading report by changing critical facts while maintaining coherent context and a journalistic tone.

Table 1: Representative prompt templates for label-preserving and label-flipping generator agents ($\mathbf{L}_0$ to $\mathbf{L}_3$ representative LLM-generated fake news settings with progressively increasing modification intensity).

2023). We therefore incorporate role-based persona conditioning into Π_{LP} and Π_{LF}, which guides generation toward professional news rewriting or controlled content modification (e.g., "You are a news editor, ...”). Representative prompt examples are provided in Table 1.

Alignment Gate. To ensure that adversarial samples remain semantically meaningful and functionally aligned with their intended generation strategies, we equip all generator agents with an alignment gate (AG) that performs hard quality filtering based on contextual, semantic, and factual consistency between the generated text $\tilde{x}$ and the source content x.

For each source–candidate pair $(x, \tilde{x})$, AG computes three normalized consistency scores in $[0, 1]$. Contextual consistency $s^{ctx}(x, \tilde{x})$ is measured using BERTScore (Zhang et al., 2019), which evaluates topical and discourse-level alignment to prevent topic drift. Semantic consistency $s^{sem}(x, \tilde{x})$ is computed as the cosine similarity between sentence-level representations, enabling detection of semantic incoherence even when surface context appears relevant. Factual consistency $s^{fact}(x, \tilde{x})$ is measured using AlignScore (Zha et al., 2023), which is designed to detect fine-grained factual misalignment and extrinsic hallucinations.

To avoid heuristic threshold selection, we maintain an unfiltered generation cache $\mathcal{C}$ that stores all generated candidates together with their consistency scores. Contextual and semantic alignment thresholds are obtained using low-tail quantiles: $\tau_{ctx} = Q_\alpha(\{s^{ctx}\}_{\mathcal{C}})$, $\tau_{sem} = Q_\alpha(\{s^{sem}\}_{\mathcal{C}})$, where $Q_\alpha(\cdot)$ denotes the α-quantile. These thresholds enforce a minimum level of topical and semantic alignment for all retained samples.

After filtering candidates that satisfy both contextual and semantic constraints, factual consistency thresholds are estimated in a strategy-aware manner. Specifically, for label-preserving samples, the factual threshold is defined as: $\tau_{fact}^{\text{LP}} = Q_\alpha(\{s^{fact}\}_{\mathcal{C}'_{\text{LP}}})$, which enforces high factual fidelity to preserve label consistency. For label-flipping samples, the factual threshold is computed as: $\tau_{fact}^{\text{LF}} = Q_\alpha(\{s^{fact}\}_{\mathcal{C}'_{\text{LF}}})$, allowing controlled factual deviation while avoiding trivial rewrites or uncontrolled hallucinations.

Retention Rule. A generated sample $\tilde{x}$ is retained if it satisfies the contextual and semantic thresholds, and additionally meets the strategy-specific factual constraint. Only retained samples are inserted into the adversarial buffer for detector training.

3.4 Hybrid Multi-Agent Detector (FNC-D)

We design FNC-D as a hybrid multi-agent detector that separates high-level semantic analysis from final judgment, enabling LLMs to focus on structured reasoning while delegating reliable classification to a SLM. The detailed design of FNC-D is described as follows.

Dimensional Division. A single LLM often struggles to analyze news from diverse perspectives and effectively integrate insights from multiple angles. FNC-D addresses this limitation by decomposing fake news detection into three complementary dimensions, $\mathcal{P} = \{\mathcal{P}_i\}_{i=1}^{3}$, namely, factual consistency, semantic plausibility, and linguistic style, which together capture the core perspectives required for assessing news veracity.

For each dimension $\mathcal{P}_i$, FNC-D assigns a dedicated group of n LLM-based reviewer agents, denoted as $\mathcal{A}_{\mathcal{P}_i} = \{\mathcal{A}_{\mathcal{P}_i}^j \mid j = 1, \ldots, n\}$, where agents anchor their analysis to the same evaluation dimension while adopting heterogeneous reasoning strategies $\mathcal{R} = \{\mathcal{R}_j \mid j = 1, ..., n\}$ to examine the input text $\tilde{x}$, thereby mitigating reasoning homogeneity and reducing individual model bias. The agents then engage in structured multi-round deliberation to reconcile disagreements and produce

robust dimension-level assessments, the design of which is detailed below.

Structured Multi-Round Debate and Integration. In the first round $t = 0$, each agent $\mathcal{A}^j_{\mathcal{P}_i} \in \mathcal{A}_{\mathcal{P}_i}$ independently generates a dimension-specific judgment under its assigned reasoning strategy $\mathcal{R}_j$:

$$\mathbf{re}^0_{\mathcal{P}_i,j} = \mathcal{A}^j_{\mathcal{P}_i}(\tilde{x} \mid \mathcal{R}_j), \tag{1}$$

where $\mathbf{re}^0_{\mathcal{P}_i,j}$ denotes the corresponding judgment for dimension $\mathcal{P}_i$.

Then, at each subsequent round $t > 0$, agents exchange their judgments from the round $t-1$. For agent $\mathcal{A}^j_{\mathcal{P}_i}$, the debate history is updated as

$$h^{t-1}_{\mathcal{P}_i,j} = \Big[(\tilde{x}, \mathbf{re}^{t-1)}_{\mathcal{P}_i,j}), \{(\tilde{x}, \mathbf{re}^{t-1}_{\mathcal{P}_i,k})\}_{k\neq j}\Big]. \tag{2}$$

Conditioned on the updated history, each agent revises its judgment in the next round:

$$\mathbf{re}^t_{\mathcal{P}_i,j} = \mathcal{A}^j_{\mathcal{P}_i}(\tilde{x} \mid h^{t-1}_{\mathcal{P}_i,j}, \mathcal{R}_j). \tag{3}$$

This iterative process allows agents to identify inconsistencies, incorporate alternative perspectives, and gradually refine their judgments.

After each round t, we check whether all agents reach an agreement, i.e., $\mathbf{re}^t_{\mathcal{P}_i,1} = \mathbf{re}^t_{\mathcal{P}_i,2} = \cdots = \mathbf{re}^t_{\mathcal{P}_i,n}$.

If consensus is achieved, the agreed judgment is directly taken as the final assessment for dimension $\mathcal{P}_i$. Otherwise, when the maximum number of rounds T is reached, the final result is determined by majority voting:

$$\mathbf{S}_{\mathcal{P}_i} = \text{mode}\left(\{\mathbf{re}^T_{\mathcal{P}_i,j}\}^n_{j=1}\right), \tag{4}$$

where $\mathbf{S}_{\mathcal{P}_i}$ denotes the dimension-level assessment for $\mathcal{P}_i$.

Next, a secretary agent is introduced to mediate between heterogeneous LLM-based deliberation and stable downstream classification. It aggregates the dimension-specific reasoning outputs, removes redundancy, and reformats them into a unified and structured representation. The resulting rationale sequence $\mathbf{S}(\tilde{x}) = \{\mathbf{S}_{\mathcal{P}_i}\}^3_{i=1}$ provides a concise summary of multi-dimensional analyses and is subsequently fed into the SLM-based classifier for final decision making.

SLM-based Classifier. To ground multi-dimensional rationale in the original news content, FNC-D employs a stable SLM-based classifier that jointly models the structured rationale $\mathbf{S}(\tilde{x})$ and the input text $\tilde{x}$. Both $\tilde{x}$ and $\mathbf{S}(\tilde{x})$ are encoded using a shared encoder (BERT in our work) to obtain contextualized token representations, $\mathbf{H}_{\tilde{x}} = \text{BERT}(\tilde{x})$ and $\mathbf{H}_{\mathbf{S}} = \text{BERT}(\mathbf{S}(\tilde{x}))$.

To explicitly align dimension-level reasoning with corresponding content in the news, we apply a cross-attention mechanism from the rationale representations to the input text representations:

$$\text{Attn}(\mathbf{H}_{\mathbf{S}}, \mathbf{H}_{\tilde{x}}) = \text{softmax}\left(\frac{(\mathbf{H}_{\mathbf{S}}W_Q)(\mathbf{H}_{\tilde{x}}W_K)^\top}{\sqrt{d}}\right)\mathbf{H}_{\tilde{x}}W_V. \tag{5}$$

The attended features are then fused via a residual connection and layer normalization,

$$\tilde{\mathbf{H}} = \text{LayerNorm}\big(\mathbf{H}_{\mathbf{S}} + \text{Attn}(\mathbf{H}_{\mathbf{S}}, \mathbf{H}_{\tilde{x}})\big). \tag{6}$$

Next, the fused representation is aggregated by mean pooling, i.e., $\mathbf{v} = \text{Pool}(\tilde{\mathbf{H}}) \in \mathbb{R}^d$, and passed to a lightweight MLP to produce the final prediction $\hat{\mathbf{p}}$:

$$\hat{\mathbf{p}} = \text{softmax}\big(W_2\,\sigma(W_1\mathbf{v} + b_1) + b_2\big), \tag{7}$$

where $\sigma(\cdot)$ denotes a nonlinear activation function.

3.5 Adversarial Training

FNC is trained under an adversarial learning paradigm. Given an input sample $\tilde{x}$ with ground-truth label $\tilde{y}$, the detector is optimized to correctly classify adversarially generated news using the standard cross-entropy loss:

$$\mathcal{L}_{\text{det}}(\tilde{x}, \tilde{y}) = -\log \hat{\mathbf{p}}_{\tilde{y}}. \tag{8}$$

In parallel, the generator is updated via a prompt-based optimization strategy rather than direct parameter learning. Inspired by prompt optimization methods such as OPRO (Zhang et al., 2024), we employ an automated prompt engineer to iteratively refine the generation prompts based on detection feedback. At iteration r, the current prompt set Π^r, together with the detector's prediction scores $\hat{\mathbf{p}}^r$, is provided to the LLM-based prompt engineer to produce an updated prompt set:

$$\Pi^{r+1} = \text{LLM}\big(\Pi^r, \hat{\mathbf{p}}^r\big). \tag{9}$$

This feedback-driven refinement enables the generator to progressively produce more challenging and diverse adversarial samples, reducing the risk that the detector simply adapts to static stylistic patterns from a fixed generation process.

To maintain consistent quality control as generation strategies evolve, the alignment thresholds

in AG ($\tau_{ctx}, \tau_{sem}, \tau_{fact}^{\text{LP}}, \tau_{fact}^{\text{LF}}$) are periodically re-estimated from the empirical score distributions accumulated in the generation cache. As prompt refinement shifts the distribution of generated samples, these thresholds adapt accordingly, ensuring that retained adversarial examples remain semantically valid and functionally aligned with their intended training roles.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate our method on three widely used fake news detection datasets: CoAID (Cui and Lee, 2020) and the FakeNewsNet benchmark (Shu et al., 2020), which comprises the PolitiFact and GossipCop datasets. All datasets are strictly divided into training, validation, and test subsets, and all final evaluations are conducted on previously unseen samples.

The LLM-generated subsets include both controllable adversarial samples used during adversarial training and more open-ended fake news samples generated through direct prompts such as "Please generate a fake news article...", which are further used for robustness evaluation.

Baseline Models. We evaluate FNC against three groups of baseline methods. The first group (G1) consists of SLM-based detectors, including PET (Schick and Schütze, 2021), which reformulates textual inputs as cloze-style questions with task descriptions to detect fake news, ARG (Hu et al., 2024), a hybrid detection framework that combines LLMs with SLMs, and SheepDog (Wu et al., 2024), a fake news detection model specifically designed to defend against style-based attacks. The second group (G2) includes LLM-based detection approaches, such as the standalone LLM-based detector GPT-4o-mini (Hurst et al., 2024) and the adversarial and explainable fake news detection framework LLM-GAN (Wang et al., 2024b). The third group (G3) comprises multi-agent debate-based methods, including ChatEval (Chan et al., 2023), which reaches final decisions through discussions among multiple agents, and TED (Liu et al., 2025), which determines news veracity via structured debates between opposing agents.

Implementation Details. All experiments are conducted on a platform equipped with Intel Xeon Gold 5218 CPUs and NVIDIA A100 GPUs with 80GB memory. The framework is implemented in Python using PyTorch. All large language models are accessed via the OpenAI and DeepSeek APIs. Specifically, GPT-4o-mini (Hurst et al., 2024) and DeepSeek-v3 (Liu et al., 2024) are employed for adversarial fake news generation in FNC-G as well as for LLM-based semantic reasoning during fake news detection in FNC-D, while the final decision is produced by the BERT-based classifier. Moreover, the α-quantile in the alignment gate is set to $\alpha = 0.05$. We set the number of agents in each dimension-specific group to $n = 3$ and the maximum number of rounds $T = 3$, covering representative reasoning strategies including Chain-of-Thought (CoT), Self-Backtracking Prompting (SBP), and Program-of-Thought (PoT), which provides a balanced trade-off between robustness, efficiency, and deployment cost.

4.2 Experiments Results

4.2.1 Robustness Against LLM-Generated Datasets

As reported in Table 2, we conduct comprehensive comparisons across both human-written and LLM-generated datasets, against representative baselines. Specifically, the detector based on DeepSeek is referred to as FNC (DS), while the detector based on GPT is denoted as FNC (GPT).

Effectiveness of the Detector FNC-D on Human-written Datasets. On human-written datasets, FNC (DS/GPT) consistently achieves competitive performance in most settings. Although it slightly underperforms TED on PolitiFact, FNC (DS/GPT) attains the better results on the larger GossipCop and CoAID datasets. Moreover, the progressive improvement observed from G1 to G3 indicates the clear benefit of multi-agent deliberation, while FNC further strengthens this advantage through structured, dimension-aware collaboration.

Robustness of FNC-D on LLM-Generated Datasets. Despite the greater difficulty of detecting LLM-generated fake news, FNC consistently demonstrates strong robustness. Notably, on the PolitiFact dataset, FNC (DS) achieves a detection accuracy on LLM-generated samples that is 3.37% higher than on the corresponding human-written subset. On the other two datasets, although the detection performance of the dataset generated by the LLM has decreased compared to the human-written dataset, the decline, approximately 10%, is considerably smaller than that observed in other baseline methods, which experience a reduction of approxi-

Group	Method	PolitiFact				GossipCop				CoAID			
		Human-Written		LLM-Generated		Human-Written		LLM-Generated		Human-Written		LLM-Generated	
		Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
G1	PET	85.56	85.51	74.49	73.23	74.75	74.63	70.68	70.57	83.51	82.66	63.81	62.50
	DeepSeek	86.36	85.70	83.75	81.12	85.00	84.39	75.30	73.69	79.63	79.50	70.25	69.80
	ARG	88.90	87.29	79.65	79.60	87.80	79.00	75.61	71.59	85.73	84.67	73.45	71.67
	SheepDog	88.44	88.39	80.99	78.50	75.77	53.54	70.63	69.30	76.86	76.80	68.97	66.53
G2	GPT-4o-mini	85.89	84.31	55.40	53.63	86.30	68.79	62.21	51.90	87.93	86.71	61.57	59.20
	LLM-GAN	89.67	89.01	73.33	72.60	89.60	82.35	63.86	70.91	88.73	87.20	72.60	72.13
G3	ChatEval	90.26	90.20	78.57	77.79	86.97	73.85	69.56	68.79	83.45	83.28	68.50	67.15
	TED	**91.73**	89.60	83.50	82.13	89.21	80.35	79.21	79.20	87.53	87.64	75.73	74.29
Ours	**FNC(DS)**	90.47	**90.40**	**93.84**	**92.59**	87.92	87.34	82.67	81.09	92.53	90.57	80.37	80.26
	FNC(GPT)	89.63	88.90	87.83	87.35	**89.65**	**88.50**	**84.56**	**82.21**	**92.68**	**91.53**	**83.36**	**82.58**

Table 2: Performance comparison of FNC(DS) and FNC(GPT) with baseline methods on paired human-written and LLM-generated news datasets.

Method	PolitiFact		GossipCop		CoAID	
	H	L	H	L	H	L
BERT	85.22	72.31	74.60	68.98	73.13	53.97
FNC-BERT	**90.47**	**93.84**	87.92	82.67	**92.53**	80.37
DeBERTa	86.33	75.10	73.86	70.80	74.59	64.71
FNC-DeBERTa	89.12	86.79	**88.56**	**83.79**	89.17	**80.47**

Table 3: Adaptability of the FNC to different SLM backbones on human-written (H) and LLM-generated (L) news subsets (Acc.%).

Method	PolitiFact		GossipCop		CoAID	
	H	L	H	L	H	L
FNC	**90.47**	**93.84**	**87.92**	**82.67**	**92.53**	**80.37**
w/o AG	88.51	87.53	84.65	79.27	83.71	77.89
w/o MA	85.22	72.31	74.60	68.98	73.13	53.97
w/o SLM	87.53	84.79	82.69	80.52	82.13	76.7

Table 4: Component-wise ablation study on human-written (H) and LLM-generated (L) news subsets (Acc.%).

mately 15%. Despite this context, FNC (DS/GPT) still outperforms competing baselines by an average margin of approximately 6%. Compared with controllable rewriting-based adversarial samples used during training, the evaluation subsets additionally contain more diverse LLM-generated fake news patterns.

These results indicate that FNC does not primarily rely on surface-level cues specific to human-written text, but instead learns more generalizable decision criteria that transfer effectively to LLM-generated content. Overall, this confirms the robustness of FNC under distribution shift and its practical effectiveness in detecting LLM-generated fake news.

4.2.2 Adaptability to SLM / LLM Backbones

To evaluate the adaptability of FNC across different model backbones, we instantiate FNC with two widely used LLMs, GPT-4o-mini (GPT) and DeepSeek-v3 (DS), as reported in Table 2, and two representative SLMs, BERT (Koroteev, 2021) and DeBERTa (He et al., 2020), as reported in Table 3. When varying the SLM backbone, the LLM component of FNC is fixed to DeepSeek for fair comparison, whereas when varying the LLM backbone, FNC is instantiated with BERT as the SLM.

As shown in Table 2, FNC(DS) consistently outperforms the standalone DS model across all datasets, with particularly pronounced improvements on LLM-generated data. For instance, on the PolitiFact-based LLM-generated dataset, FNC(DS) achieves approximately a 10% accuracy gain over the single DS model. A similar trend is observed for the GPT-based setting, where FNC(GPT) consistently surpasses the standalone GPT across all datasets.

Furthermore, Table 3 shows that FNC consistently outperforms the corresponding standalone SLMs across all backbones, demonstrating that the proposed framework generalizes well to different SLM choices. Comparing different SLM backbones within FNC, FNC-DeBERTa achieves about 1%–2% higher accuracy on larger datasets such as GossipCop and CoAID, whereas FNC-BERT performs slightly better on the smaller PolitiFact dataset.

4.2.3 Ablation Study

Key Components Ablation. We conduct ablation studies on the key components of FNC, including the alignment gate (AG), the multi-dimensional multi-agent deliberation module (MA), and the SLM-based classifier (SLM).

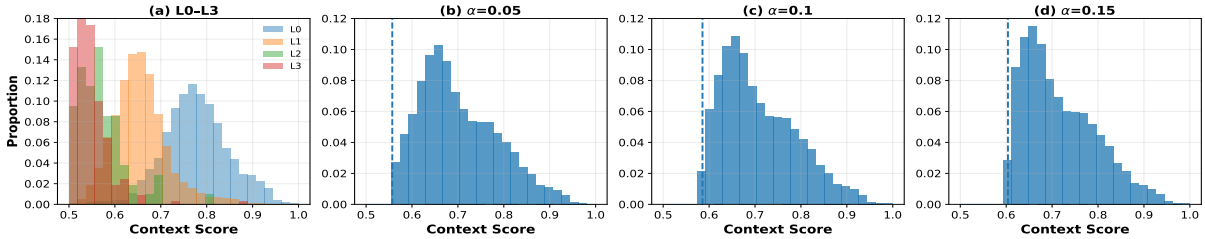

Figure 3: Contextual consistency analysis on CoAID. From left to right, the figure shows distributions under different prompt regimes ($\mathbf{L}_0$–$\mathbf{L}_3$), followed by the effects of quantile-based filtering with $\alpha = 0.05$, 0.10, and 0.15.

Method	PolitiFact		GossipCop		CoAID	
	H	L	H	L	H	L
FNC	**90.47**	**93.84**	**87.92**	**82.67**	**92.53**	**80.37**
all COT	86.79	84.31	83.50	78.53	85.33	72.68
all SBP	83.78	79.60	82.79	75.32	81.96	67.31
all POT	85.61	83.57	83.97	79.67	87.67	68.59

Table 5: Ablation study of reasoning strategies on human-written (H) and LLM-generated (L) news subsets (Acc.%).

As shown in Table 4, removing the AG leads to degraded adversarial sample quality due to increased semantic drift and factual inconsistency, resulting in performance drops of less than 5% for both BERT- and DeepSeek-based FNC. In contrast, removing the multi-agent (MA) module causes the most severe degradation on LLM-generated datasets. For example, on the GossipCop-based LLM-generated data, detection accuracy drops sharply from 82.67% to 68.98%, as the detector degenerates into a single fine-tuned SLM. This highlights the critical role of structured multi-agent deliberation in handling LLM-generated fake news. Finally, removing the SLM classifier and relying solely on agent deliberation also leads to performance degradation, albeit to a lesser extent, indicating that the SLM head provides additional stability and reliability for final decisions.

The above ablation results demonstrate that all components are necessary, with the multi-agent deliberation module being the most influential factor for robust detection.

Reasoning Strategy Ablation. To analyze the contribution of different reasoning strategies, we compare the full FNC with three degraded variants that rely on a single reasoning paradigm.

As shown in Table 5, the complete FNC that integrates multiple reasoning strategies consistently outperforms its variants relying on a single reasoning paradigm. This performance gap highlights the limitation of relying on a single reasoning strategy and demonstrates the necessity of combining complementary reasoning paradigms to robustly handle diverse LLM-generated fake news.

Hyperparameter Analysis. Figure 3 provides a distributional analysis of contextual consistency on CoAID. Different prompt regimes ($\mathbf{L}_0$–$\mathbf{L}_3$) induce clearly separable and ordered distributions, validating the necessity of explicitly distinguishing prompt severities and label-preserving versus label-flipping generation. Quantile-based filtering with $\alpha = 0.05$ removes only extreme low-alignment outliers while largely preserving the overall distributional structure. In contrast, larger α values lead to stronger truncation and noticeable contraction of the distributions, which would suppress moderately informative samples. These observations support $\alpha = 0.05$ as a balanced choice for quality control without sacrificing adversarial diversity.

5 Conclusion

In this work, we propose FNC, a robust multi-agent adversarial framework for detecting fake news generated by LLM. FNC adapts to evolving generation strategies through adversarial co-evolution between a generator and a detector. On the generation side, an alignment gate enforces contextual, semantic, and factual consistency to ensure high-quality adversarial samples. On the detection side, FNC integrates multi-agent LLM-based reasoning with stable SLM-based discrimination, decomposing verification into complementary dimensions and mitigating hallucination and instability of single-LLM judgments through collaborative deliberation and evidence alignment. Experimental results across multiple datasets demonstrate that FNC achieves stable and robust performance in detecting LLM-generated fake news.

6 Limitations

We acknowledge several limitations of this work that suggest directions for future research. First, although our experiments show that diverse prompt designs can effectively increase the diversity and difficulty of generated adversarial samples, we do not perform a systematic analysis of prompt sensitivity within each prompt category. In particular, the impact of different prompt formulations or prompt strengths on detection performance and robustness has not been thoroughly examined. Second, our framework adopts fixed hyperparameter settings with the number of agents per dimension $n = 3$ and the number of debate rounds $T = 3$, which provide a favorable trade-off between performance and computational cost. In practice, both the number of agents and debate rounds can be flexibly adjusted according to different resource budgets and deployment requirements. However, due to resource constraints, we do not explore the performance upper bounds under larger n or T. Exploring more efficient multi-agent collaboration strategies and adaptive debate mechanisms remains an important direction for future work. Third, our implementation relies on only two mainstream LLMs under limited cost constraints. Incorporating a broader and more diverse set of LLMs could further enrich adversarial sample diversity and potentially improve detector generalization, which we leave for future investigation.

References

Muhammad Zain Ali, Yuxia Wang, Bernhard Pfahringer, and Tony C Smith. 2025. Detection of human and machine-authored fake news in urdu. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3419–3428.

Amrita Bhattacharjee and Huan Liu. 2024. Fighting fire with fire: can chatgpt detect ai-generated text? *ACM SIGKDD Explorations Newsletter*, 25(2):14–21.

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. 2023. Chateval: Towards better llm-based evaluators through multi-agent debate. *arXiv preprint arXiv:2308.07201*.

Canyu Chen and Kai Shu. 2024. Can llm-generated misinformation be detected? In *The Twelfth International Conference on Learning Representations (ICLR)*.

Michał Choraś, Konstantinos Demestichas, Agata Giełczyk, Álvaro Herrero, Paweł Ksieniewicz, Konstantina Remoundou, Daniel Urda, and Michał Woźniak. 2021. Advanced machine learning techniques for fake news (online disinformation) detection: A systematic mapping study. *Applied Soft Computing*, 101:107050.

Limeng Cui and Dongwon Lee. 2020. Coaid: Covid-19 healthcare misinformation dataset. *arXiv preprint arXiv:2006.00885*.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

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.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Ashwaq Khalil, Moath Jarrah, Monther Aldwairi, and Yaser Jararweh. 2021. Detecting arabic fake news using machine learning. In *2021 second international conference on intelligent data science technologies and applications (IDSTA)*, pages 171–177. IEEE.

Mikhail V Koroteev. 2021. Bert: a review of applications in natural language processing and understanding. *arXiv preprint arXiv:2103.11943*.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and 1 others. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Yuhan Liu, Yuxuan Liu, Xiaoqing Zhang, Xiuying Chen, and Rui Yan. 2025. The truth becomes clearer through debate! multi-agent systems with large language models unmask fake news. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 504–514.

Paula Maddigan and Teo Susnjak. 2023. Chat2vis: Generating data visualizations via natural language using chatgpt, codex and gpt-3 large language models. *Ieee Access*, 11:45181–45193.

Shiwen Ni, Jiawen Li, and Hung-Yu Kao. 2021. Mvan: Multi-view attention networks for fake news detection on social media. *IEEE Access*, 9:106907–106917.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Md Shohel Rana, Mohammad Nur Nobi, Beddhu Murali, and Andrew H Sung. 2022. Deepfake detection: A systematic literature review. *IEEE access*, 10:25494–25513.

Dorsaf Sallami, Yuan-Chen Chang, and Esma Aïmeur. 2024. From deception to detection: The dual roles of large language models in fake news. *arXiv preprint arXiv:2409.17416*.

Timo Schick and Hinrich Schütze. 2021. Exploiting cloze-questions for few-shot text classification and natural language inference. In *Proceedings of the 16th conference of the European chapter of the association for computational linguistics: main volume*, pages 255–269.

Ali Shafahi, Mahyar Najibi, Mohammad Amin Ghiasi, Zheng Xu, John Dickerson, Christoph Studer, Larry S Davis, Gavin Taylor, and Tom Goldstein. 2019. Adversarial training for free! *Advances in neural information processing systems*, 32.

Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. 2020. Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. *Big data*, 8(3):171–188.

Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake news detection on social media: A data mining perspective. *ACM SIGKDD explorations newsletter*, 19(1):22–36.

Giovanni Spitale, Nikola Biller-Andorno, and Federico Germani. 2023. Ai model gpt-3 (dis)informs us better than humans. *Science Advances*, 9(26):eadh1850.

Chong Tian, Qirong Ho, and Xiuying Chen. 2025. A symbolic adversarial learning framework for evolving fake news generation and detection. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 12307–12321.

Lionel Z. Wang, Yiming Ma, Renfei Gao, Beichen Guo, Han Zhu, Wenqi Fan, Zexin Lu, and Ka Chung Ng. 2024a. Megafake: A theory-driven dataset of fake news generated by large language models. *arXiv preprint arXiv:2408.11871*.

Xinyu Wang, Wenbo Zhang, Sai Koneru, Hangzhi Guo, Bonam Mingole, S Shyam Sundar, Sarah Rajtmajer, and Amulya Yadav. 2025. Have llms reopened the pandora's box of ai-generated fake news? In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2795–2811.

Yifeng Wang, Zhouhong Gu, Siwei Zhang, Suhang Zheng, Tao Wang, Tianyu Li, Hongwei Feng, and Yanghua Xiao. 2024b. Llm-gan: construct generative adversarial network through large language models for explainable fake news detection. *arXiv preprint arXiv:2409.01787*.

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*, pages 3367–3378.

Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. 2023. Alignscore: Evaluating factual consistency with a unified alignment function. *arXiv preprint arXiv:2305.16739*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Tuo Zhang, Jinyue Yuan, and Salman Avestimehr. 2024. Revisiting opro: The limitations of small-scale llms as optimizers. *arXiv preprint arXiv:2405.10276*.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, and 1 others. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 1(2).

Jiawei Zhou, Yixuan Zhang, Qianni Luo, Andrea G. Parker, and Munmun De Choudhury. 2023. Synthetic lies: Understanding ai-generated misinformation and evaluating algorithmic and human solutions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*.