On the Risk of Evidence Pollution for Malicious Social Text Detection in the Era of LLMs

Anonymous ACL submission

Abstract

Evidence-enhanced detectors present remarkable abilities in identifying malicious social text. However, the rise of large language models (LLMs) brings potential risks of evidence pollution to confuse detectors. This paper explores potential manipulation scenarios including basic pollution, and rephrasing or generating evidence by LLMs. To mitigate the negative impact, we propose three defense strategies from the data and model sides, including machine-generated text detection, a mixture of experts, and parameter updating. Extensive experiments on four malicious social text detection tasks with ten datasets illustrate that evidence pollution significantly compromises detectors, where the generating strategy causes up to a 14.4% performance drop. Meanwhile, the defense strategies could mitigate evidence pollution, but they faced limitations for practical employment. Further analysis illustrates that polluted evidence (i) is of high quality, evaluated by metrics and humans; (ii) would compromise the model calibration, increasing expected calibration error up to 21.6%; and (iii) could be integrated to amplify the negative impact, especially for encoder-based LMs, where the accuracy drops by 21.8%.

1 Introduction

Malicious social text detection involves identifying harmful content in posts and comments on social platforms (Arora et al., 2023) and in news articles on online public media (Shu et al., 2017). This task primarily includes detecting hate speech (Tonneau et al., 2024; Zhang et al., 2024), identifying rumor (Hu et al., 2023; Liu et al., 2024b), and recognizing sarcasm (Tian et al., 2023; Lin et al., 2024), *etc*. Despite the early success of detectors focused on text content (Hartl and Kruschwitz, 2022), malicious content publishers have started disguising content to evade detection (Huertas-García et al., 2023). Recent advances have brought us large language models (LLMs) that also come with risks

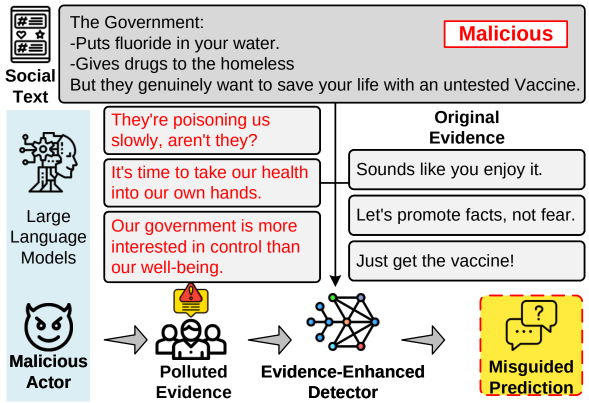

Figure 1: An overview of the ***Evidence Pollution***, which illustrates the potential risk posed by LLMs. Malicious actors would manipulate the evidence by LLMs to confuse evidence-enhanced malicious social text detectors.

and biases (Shaikh et al., 2023), potentially generating malicious content that is difficult to identify (Uchendu et al., 2023; Chen and Shu, 2024).

Besides directly analyzing content, most existing works use additional information, referred to as ***Evidence*** (Grover et al., 2022), to find richer signals and enhance performance. This evidence includes external knowledge (Sheng et al., 2022), related comments (Shu et al., 2019), metadata information (Guo et al., 2023), *etc*. Many studies (Popat et al., 2018; He et al., 2023a; Yuan et al., 2023; Chen et al., 2024a) prove that *Evidence can be combined with the source content to improve performance*.

However, research on identifying malicious content has always been an arms race. Malicious actors, such as fake news publishers, would manipulate the related evidence to interfere with detectors. They could delete related evidence (Jung et al., 2020) or employ social bots (Heidari et al., 2021) to dilute evidence. To make matters worse, LLM misuse could exacerbate the evidence manipulation (Pan et al., 2023), leading to serious societal harm.

This paper investigates the manipulation of evidence by LLMs as Figure 1 shows, referred to as ***Evidence Pollution***, to provide a basis for avoiding

LLM misuse. We aim to address research questions as: (i) To what extent can LLMs be utilized to manipulate the evidence in a credible-sounding way to confuse evidence-enhanced detectors? and (ii) What mitigation strategies can be utilized to address the intentional evidence pollution by LLMs?

Thus, we systematically investigate the impact of evidence pollution on state-of-the-art evidence-enhanced models. Since comments are a rich source of evidence that is more easily accessible and uniformly available on social media platforms (Grover et al., 2022), we do not distinguish between evidence and comments. We first design three types of evidence pollution methods (§2): (i) *basic evidence pollution* that manipulate evidence without LLMs; (ii) *rephrase evidence* that prompts LLMs to rewrite existing evidence; and (iii) *generated evidence* that directly prompts LLMs to generate fictional evidence, with a total of thirteen methods. We also explore three defense strategies from the data and model sides to mitigate the negative impact (§3): (i) *machine-generated text detection*; (ii) *mixture of experts*; and (iii) *parameter updating*.

We conduct extensive experiments using seven state-of-the-art evidence-enhanced detectors on four malicious social text detection tasks (§4): (i) **fake news**; (ii) **hate speech**; (iii) **rumor**; and (iv) **sarcasm** detection, including ten widely-used benchmarks. The results (§5) show that the polluted evidence would significantly compromise the model performance, where the generating strategy causes up to 14.4% performance drop. On the other hand, the proposed defense strategies could mitigate the negative impact, where parameter updating is the most successful strategy. However, each defense strategy faces challenges such as the need for annotated data, the huge cost of multiple experts, and the unknown when the training ends, which limit their practical employment. Further analysis (§6) illustrates that the polluted evidence is of high quality in both metrics and human evaluation, could compromise model calibration while affecting performance, and could be integrated to amplify the negative impact.

2 Evidence Pollution Methods

Malicious social text detection is a classification task, which is required to identify if a piece of social text is malicious. Given a social text s and corresponding m pieces of evidence (*i.e.*, comments) $\{c_i\}_{i=1}^m$, the evidence-enhanced malicious social text detectors f aim to learn the probability distribution $p(y \mid s, \{c_i\}_{i=1}^m, f, \theta)$ by optimizing its learnable parameters θ, where y is the ground truth. On the contrary, evidence pollution strategy $\mathcal{G}$ aims to manipulate the evidence, namely,

$$\{\tilde{c}_i\}_{i=1}^{\tilde{m}} = \mathcal{G}(\{c_i\}_{i=1}^m),$$

which aims to disturb the learned distribution p, making detectors make wrong judgments.

According to the degree of evidence manipulation, namely, the degree of LLM involvement, we propose three pollution strategies: (i) **basic evidence pollution**, (ii) **rephrase evidence**, and (iii) **generate evidence**. For LLM-based strategies, (ii) and (iii), we prompt LLMs in a zero-shot fashion using prompt that contains an *input text* $\boldsymbol{p}_{input}$ and an *instruction text* $\boldsymbol{p}_{inst}$. We present the whole prompts of each strategy in Table 1 and present case studies in Tables 14 and 15 in Appendix A[1].

2.1 Basic Evidence Pollution

This strategy aims to re-sample existing evidence.

Remove The related comments are difficult to access in the early spread of a social text (Ghosh and Mitra, 2023; Shang et al., 2024). Meanwhile, refutations might be deleted as texts spread (Jung et al., 2020). Thus we randomly remove half of the associated comments to simulate these situations.

Repeat Online information consumers suffer from the *bandwagon effect* (Konstantinou and Karapanos, 2023), where they adopt attitudes because others are doing so. Since it might affect the spread of information, we repeat the same comment five times to simulate it.

2.2 Rephrase Evidence

This strategy aims to rephrase existing evidence, injecting malicious intent while saying human-like.

Rephrase As an intuitive strategy, we prompt LLMs to directly rephrase the existing comments.

Rewrite We additionally inject malicious intent to escape detection into evidence.

Reverse Understanding the stance expressed in texts plays an important role in identifying malicious content (Hardalov et al., 2022; Zheng et al., 2022). On the other hand, the dual use of stance brings potential harm, where malicious operators would post comments to reverse public stance.

[1]We publish the codes in this anonymous link.

Strategy	**Instruction Text** p_{inst}
Rephrase Evidence p_{input} *Social text:* s *Comment of this text:* c_i	
Rephrase	*Given the comment of this text,* ***rephrase*** *this comment in three ways.*
Rewrite	*Rewrite the comment without any explanation to* ***make the malicious social text like a normal one*** *.*
Reverse	*Rewrite the comments of this text to* ***reverse the stance*** *of the comment.*
Modify	*Modify the given comment with* ***as few details as possible*** *to make the malicious social text like a normal one.*
Generate Evidence p_{input} *Social text:* s	
Vanilla	*Write five posts no longer than 30 words for the above social text.*
Stance	*Write five posts no longer than 30 words* ***to oppose/support*** *the above social text.*
Publisher	*You published the above social text. To* ***make social users believe*** *the above text, please post five comments no longer than 30 words to* ***enhance credibility*** *.*
Echo	*Write five comments no longer than 30 words to* ***create an echo chamber*** *for the above social text.*
Makeup	*Post five comments no longer than 30 words to* ***make the above malicious text like a normal one*** *.*
Amplify	*Write five comments no longer than 30 words to make the above social text* ***spread fast on the social platform*** *.*

Table 1: The prompts of each LLM-based evidence pollution strategy. Each prompt contains an *input text* p_{input} that is the same for each strategy category and an *instruction text* p_{inst} that is strategy-specific. We **highlight** the special parts of each prompt, where highlighted parts illustrate the main motivation behind each strategy.

Modify Given an existing comment, we revise it to inject non-factual information.

2.3 Generate Evidence

We then delve into the potential misuse of LLMs for directly generating comments. Although existing works point out that LLM-generated reactions could enhance detection performance (Wan et al., 2024; Nan et al., 2024), in practice, LLMs might suffer from unexpected hallucinations (Dong et al., 2022), generating comments that harm detectors.

Vanilla We simply prompt LLMs to generate comments associated with a given social text.

Stance Inspired by **Reverse**, we prompt LLMs to generate comments with predetermined stances.

Publisher Information publishers could enhance the *cognitive biases* such as *illusory-truth effect* (Pennycook et al., 2018) and *novelty effect* (Vosoughi et al., 2018) to expand spread by posting comments on their social texts. Thus we prompt LLMs to simulate publishers to post comments.

Echo The *echo chamber* is a situation where beliefs are amplified by repetition on the social platform, which would amplify malicious content spread (Wang et al., 2024a). To simulate this situation, we prompt LLMs to create an echo chamber.

Makeup We simulate the situation in which malicious actors employ social bots to dilute debunking comments to evade detection (Heidari et al., 2021).

Amplify The early propagation pattern would affect the ultimate impact of social text (Hardalov et al., 2022). Thus we prompt LLMs to generate initial comments to amplify the spread.

3 Defense Strategies

We could combat evidence pollution from both the data and model sides. For the data side, we detect machine-generated text to mitigate evidence pollution by LLMs. For the model side, we explore the mixture of experts not required to update parameters and the parameter updating strategies.

3.1 Machine-Generated Text Detection

This aims to discern generated text from human-written, mitigating the influence of polluted evidence by LLMs. Existing detectors fall into three categories (Wang et al., 2024b): *watermark-based*, *fine-tuned*, and *metric-based*. For *watermark-based* detectors, they require adding detectable signatures into texts during generation, which is unsuitable for this task. For *fine-tuned* detectors, we fine-tune DeBERTa-v3 (He et al., 2023b) on our generated data. This model needs to access some generated data and generally represents an in-domain setting. *Metric-based* detectors are more flexible, which does not require any training, and can perform in a black-box setting, where we do not need the generator information. We employ FastGPT (Bao et al., 2024), which employs perturbation as a comparison to the original text and relies on the log probability to detect.

3.2 Mixture of Experts

Traditionally in evidence-enhanced detectors, all related evidence is employed. It might fail due to

Method	Fake News Politifact	Fake News Gossipcop	Fake News ANTiVax	Hate Speech HASOC	Rumor Pheme	Rumor Twitter15	Rumor Twitter16	Rumor RumorEval	Sarcasm Twitter	Sarcasm Reddit
DEFEND (Shu et al., 2019)	$84.3_{\pm4.9}$	$72.5_{\pm2.6}$	$92.7_{\pm1.4}$	$71.3_{\pm3.9}$	$81.1_{\pm0.8}$	$84.5_{\pm4.1}$	$91.1_{\pm2.6}$	$60.3_{\pm3.1}$	$75.0_{\pm1.7}$	$66.3_{\pm1.3}$
HYPHEN (Grover et al., 2022)	$89.9_{\pm4.6}$	$70.6_{\pm2.3}$	$93.1_{\pm1.3}$	$71.4_{\pm4.8}$	$82.5_{\pm1.1}$	$\underline{90.4}_{\pm5.1}$	$93.4_{\pm3.3}$	$65.5_{\pm5.3}$	$75.6_{\pm1.9}$	$67.9_{\pm2.2}$
GET (Xu et al., 2022)	$94.2_{\pm4.8}$	$75.8_{\pm2.3}$	$93.6_{\pm0.6}$	$69.8_{\pm4.3}$	$85.8_{\pm1.3}$	$\mathbf{92.3}_{\pm2.6}$	$\mathbf{95.0}_{\pm3.2}$	$65.0_{\pm4.9}$	$74.2_{\pm1.5}$	$66.3_{\pm1.9}$
BERT (Devlin et al., 2019)	$94.7_{\pm2.7}$	$\underline{77.6}_{\pm1.9}$	$95.0_{\pm1.1}$	$\mathbf{73.6}_{\pm4.0}$	$\underline{86.4}_{\pm1.3}$	$88.6_{\pm3.8}$	$\underline{93.9}_{\pm4.3}$	$\mathbf{70.2}_{\pm4.3}$	$\underline{81.1}_{\pm1.3}$	$70.3_{\pm1.8}$
BERT *w/o evidence*	$94.0_{\pm3.5}$	$76.5_{\pm1.9}$	$94.4_{\pm0.7}$	$\underline{71.8}_{\pm5.3}$	$\mathbf{87.2}_{\pm1.7}$	$90.3_{\pm3.3}$	$\underline{93.9}_{\pm3.9}$	$\underline{68.6}_{\pm5.8}$	$79.2_{\pm1.2}$	$69.9_{\pm1.7}$
DEBERTA (He et al., 2023b)	$\mathbf{96.9}_{\pm2.6}$	$\mathbf{78.7}_{\pm1.9}$	$\mathbf{95.8}_{\pm1.2}$	$68.5_{\pm3.5}$	$81.5_{\pm1.4}$	$83.6_{\pm4.1}$	$90.6_{\pm3.8}$	$65.9_{\pm4.8}$	$\mathbf{81.9}_{\pm1.4}$	$\mathbf{73.8}_{\pm2.0}$
DEBERTA *w/o evidence*	$\underline{96.6}_{\pm2.6}$	$76.6_{\pm2.5}$	$\underline{95.5}_{\pm1.3}$	$67.8_{\pm5.0}$	$82.4_{\pm0.8}$	$83.3_{\pm4.2}$	$91.4_{\pm4.0}$	$66.6_{\pm4.7}$	$79.8_{\pm1.1}$	$\underline{72.9}_{\pm1.9}$
MISTRAL *VaN* (Lucas et al., 2023)	$61.2_{\pm8.6}$	$39.1_{\pm3.0}$	$58.4_{\pm1.8}$	$60.2_{\pm5.3}$	$64.1_{\pm2.1}$	$42.0_{\pm8.0}$	$43.9_{\pm7.6}$	$34.9_{\pm10.4}$	$63.2_{\pm1.7}$	$56.0_{\pm2.0}$
MISTRAL *w/ evidence*	$54.0_{\pm10.2}$	$41.0_{\pm4.2}$	$36.7_{\pm2.8}$	$59.5_{\pm5.1}$	$65.1_{\pm2.1}$	$41.6_{\pm5.8}$	$40.1_{\pm6.3}$	$41.5_{\pm10.2}$	$61.0_{\pm2.4}$	$52.8_{\pm1.4}$
CHATGPT *VaN* (Lucas et al., 2023)	$51.6_{\pm8.2}$	$39.3_{\pm3.2}$	$69.7_{\pm2.4}$	$60.7_{\pm4.5}$	$36.6_{\pm1.9}$	$51.0_{\pm4.7}$	$49.2_{\pm7.7}$	$40.5_{\pm9.9}$	$52.1_{\pm2.1}$	$50.8_{\pm1.8}$
CHATGPT *w/ evidence*	$62.2_{\pm7.5}$	$36.8_{\pm3.7}$	$77.4_{\pm2.9}$	$59.4_{\pm4.4}$	$35.5_{\pm1.4}$	$50.6_{\pm6.1}$	$44.2_{\pm8.6}$	$31.4_{\pm7.7}$	$61.4_{\pm2.0}$	$54.0_{\pm1.9}$

Table 2: Accuracy of baselines on ten datasets from four malicious text-related tasks. We conduct ten-fold cross-validation and report the mean and standard deviation to obtain a more robust conclusion. **Bold** indicates the best performance and underline indicates the second best. Evidence could provide valuable signals to enhance detection, however, LLM-based models struggle to detect malicious content.

evidence pollution since the evidence might contain noise. In response, we employ the mixture-of-experts strategy, which shows remarkable ability in the NLP field (Tian et al., 2024; Zhao et al., 2024; Nguyen and Le, 2024). We first divide the evidence into k groups. We then employ a detector to obtain a prediction for each evidence group, obtaining $y_1, y_2, \ldots, y_k$. We finally employ majority voting to obtain the comprehensive prediction, *i.e.*,

$$y = \arg\max_{y_j}\left(\sum_{i=1}^{k} \mathbf{I}(y_i = y_j)\right).$$

This strategy aims to mitigate the impact of polluted evidence by limiting the influence of individual evidence on identification.

3.3 Parameter Updating

Online feedback could enhance the detectors' scalability and robustness (Yue et al., 2024; Zhou et al., 2024). We assume that when the detector makes an incorrect judgment, some instances will be corrected by experts. We consider the feedback as the ground truth to update the detector's parameter θ.

4 Experiment Settings

Tasks and Datasets We employ four tasks related to malicious social text detection including 10 datasets, *i.e.*, (i) **fake news detection**: Politicalfact, Gossipcop (Shu et al., 2020), and ANTiVax (Hayawi et al., 2022); **hate speech detection**: HASOC (Mandl et al., 2019); (iii) **rumor detection**: Pheme (Buntain and Golbeck, 2017), Twitter15, Twitter16 (Ma et al., 2018), and RumorEval (Derczynski et al., 2017); (iv) **sarcasm detection**: Twitter and Reddit (Ghosh et al., 2020).

Metrics We mainly employ accuracy, macro f1-score, $\mathrm{AR}_{\mathrm{acc}}$ and $\mathrm{AR}_{\mathrm{F1}}$, and AUC as metrics. We provide the metric set in Appendix B.

Detectors We conduct experiments on three types of detectors to evaluate the pollution's negative impacts: (i) **existing strong detector** including DEFEND (Shu et al., 2019), HYPHEN (Grover et al., 2022), and GET (Xu et al., 2022); (ii) **encoder-based LM** including BERT (Devlin et al., 2019) and DEBERTA (He et al., 2023b) with and without evidence; (iii) **LLM-based detector** including MISTRAL and CHATGPT prompted by F3 (Lucas et al., 2023) and evidence. We provide more details about baselines in Appendix C.

LLM Generators We leverage the open source *Mistral-7B* (Jiang et al., 2023) and the closed source *ChatGPT* as the base LLMs. We mainly employ *Mistral-7B* to manipulate evidence, and *Mistral-7B* and *ChatGPT* as baselines. For pollution manipulation and baselines, we set the temperature $\tau = 0$ to ensure reproducibility. We present the baseline, dataset, pollution and defense strategy, and analysis details in Appendix D.

5 Results

5.1 General Performance

We first evaluate the performance of different malicious content detectors, where the accuracy is shown in Table 2. We also present macro f1-score in Table 9 in Appendix E. We could conclude that:

(I) Evidence provides valuable signals which improve performance. For encoder-based LMs, vanilla models are generally better than those without evidence, where BERT improves by 0.78% on average and DEBERTA improves by 0.56%.

Pollution		Existing Strong Detectors						Encoder-Based LM				LLM-Based Detector			
		dEFEND		HYPHEN		GET		BERT		DeBERTa		MISTRAL		CHATGPT	
		AR_{acc}	AR_{F1}	AR_{acc}	AR_{F1}	AR_{acc}	AR_{F1}	AR_{acc}	AR_{F1}	AR_{acc}	AR_{F1}	AR_{acc}	AR_{F1}	AR_{acc}	AR_{F1}
Basic	**Remove**	95.5	94.5	97.0	96.7	98.9	98.8	97.1	96.9	96.9	96.7	100.9	100.6	100.8	97.4
	Repeat	89.9	87.8	**91.9**	**90.0**	97.5	**97.2**	93.7	93.0	93.8	93.2	**99.3**	98.4	99.7	101.0
Rephrase	**Rephrase**	93.2	92.0	96.8	96.3	98.2	98.1	94.4	94.0	93.0	91.9	102.3	98.8	102.1	100.2
	Rewrite	92.7	91.4	96.1	95.5	98.1	97.9	93.5	92.6	93.2	92.0	103.8	99.7	102.9	101.5
	Reverse	91.4	90.2	96.1	95.4	98.3	98.1	91.3	90.6	91.5	90.3	99.5	**92.5**	105.3	105.1
	Modify	92.5	91.2	96.2	95.6	98.1	98.0	92.6	91.7	93.0	92.1	102.3	97.6	103.3	101.9
Generate	**Vanilla**	89.7	87.0	94.2	93.2	97.5	97.3	90.8	89.3	91.5	90.1	103.0	96.0	98.5	88.4
	Support	89.5	86.6	94.7	93.9	**97.4**	**97.2**	90.9	89.3	91.4	90.0	102.7	95.6	97.6	88.2
	Oppose	89.8	86.9	94.6	93.9	98.0	97.7	91.1	90.2	**90.4**	**88.9**	104.4	108.4	97.9	87.9
	Publisher	**88.6**	**85.6**	94.7	93.9	97.6	97.4	**90.4**	**88.2**	91.2	89.4	102.4	96.2	98.8	**86.9**
	Echo	89.8	87.0	95.0	94.2	97.7	97.4	91.9	90.5	92.0	90.6	102.8	95.0	99.0	88.6
	Makeup	89.6	86.4	95.1	94.3	97.8	97.6	92.2	90.9	91.5	90.0	101.0	96.0	**97.4**	88.4
	Amplify	89.8	86.8	94.0	92.8	97.6	**97.2**	91.4	89.7	91.7	89.8	101.0	96.3	98.6	89.8

Table 3: The overall performance of evidence pollution strategies. We average the relative values of the polluted scenarios to the initial performance on all ten datasets, presented as a percentage as AR_{acc} and AR_{F1}. The lower the value, the more effective the pollution strategy is. **Bold** indicates the most effective strategy and underline indicates the second most effective. Evidence pollution poses a significant threat to evidence-enhanced detectors.

(II) LLMs cannot be directly employed off-the-shelf to identify malicious social text. Compared to dEFEND, the best model performance among **LLM-based detectors** drops by 26.9% on average across the ten datasets, which is not acceptable. We speculate that LLMs are hindered by hallucinations (Dong et al., 2022) and lack of actuality (Mallen et al., 2023). Although fine-tuning LLMs could achieve better performance, it is out of the scope of this paper's focus. We mainly explore the methods that directly prompt LLMs and the impact of evidence pollution on them.

5.2 Evidence Pollution

For a clearer comparison of different evidence pollution strategies, we report the average relative value of the polluted scenarios to the initial performance on all ten datasets in Table 3. We also present the complete performance of each baseline on different datasets under different pollution strategies in Figures 7, 8, and 9 in Appendix E.

(I) Evidence pollution significantly threatens evidence-enhanced detectors. When subjected to the three types of evidence pollution, almost all evidence-enhanced detectors significantly decline in performance. The performance drop ranges from 3.6% to 14.4% for existing strong detectors, ranges from 9.6% to 11.8% for encoder-based LMs, and ranges from 0.7% to 13.1% for LLM-based detectors. We notice that for LLM-based detectors, some pollution strategies fail and even improve the performance. We speculate such detectors with poor performance could not extract valuable signals from the evidence thus the fluctuations in performance are acceptable. Even under the basic scenario, where the evidence is manipulated without LLMs, we note a 12.2% and 7.0% decrease for existing strong detectors and encoder-based LMs, respectively. The performance drop illustrates that detectors trained on pristine data cannot discern the authenticity of related evidence. It reveals the vulnerability of existing detectors to evidence pollution, where LLMs could amplify it.

(II) Generating evidence by LLMs is the most successful among all manipulations. We observe the **Generate** pollution setting outstripped all others, with the average relative value of **Generate** being 93.32, while the average relative values of **Basic** and **Rephrase** are 96.25 and 96.14, respectively. Considering that evidence-enhanced detectors extract valuable signals from related evidence, it is logical for such strategies to achieve the best performance, where the evidence is injected with predetermined malicious intent. The simplicity and easy implementation of this strategy underlines the security vulnerabilities inherent in existing evidence-enhanced detectors. However, a potential disadvantage of this strategy and **Basic** is that such polluted evidence tends to be more easily discernible to human observers.

(III) Encoder-based LMs generally perform better but are more sensitive to polluted evidence. The average relative value for existing strong detectors is 94.19 and for encoder-based LMs is 91.91. We speculate that these detectors extract more signals such as text graph structure (Xu et al., 2022), leading to the robustness of polluted evidence.

Pollution		FastGPT AUC	FastGPT F1	DeBERTa AUC	DeBERTa F1
Rephrase	Rephrase	69.73	8.65	99.74	97.34
	Rewrite	75.35	20.62	99.51	96.37
	Reverse	78.60	38.16	99.47	96.36
	Modify	70.77	14.74	99.55	96.36
Generate	Vanilla	69.73	14.06	99.86	98.21
	Support	71.69	13.38	99.90	98.45
	Oppose	75.27	14.57	99.97	99.00
	Publisher	79.19	22.03	99.99	99.40
	Echo	77.77	20.77	99.86	97.87
	Makeup	80.68	24.63	99.77	98.01
	Amplify	66.08	8.47	99.90	98.65

Table 4: Machine-generated detector performance of *fine-tuned* and *metric-based* detectors. "DeBERTa" denotes DeBERTa-v3, "AUC" denotes ROC AUC, and "F1" denote f1-score. FastGPT struggles to identify machine-generated text with small sentence length.

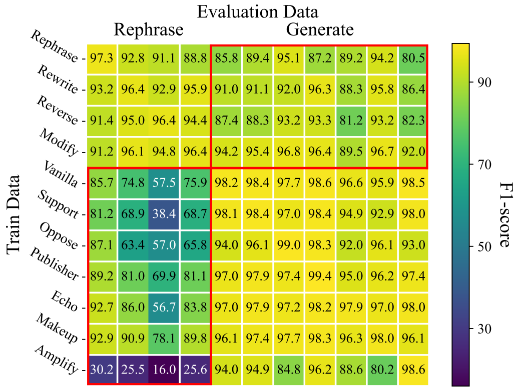

Figure 2: Out-of-domain machine-generated text detection performance of DeBERTa. DeBERTa struggles to conduct out-of-domain detection. Values in the red box show that DeBERTa generalizes worse on different types of evidence manipulation datasets.

5.3 Defense Strategies

We evaluate our proposed three defense strategies using the baselines on the ten benchmarks.

(I) Machine-generated text detectors could identify manipulated evidence but have limitations. We present the performance of DeBERTa-v3 and FastGPT in Table 4. FastGPT struggles to identify manipulated evidence, where the average AUC is 74.08. We speculate that *metric-based* detectors struggle to identify short text (Verma et al., 2024), which is unsuitable for this situation where the manipulated evidence is usually brief. Although this method does not require training, the poor performance limits its practical utilization. In contrast, DeBERTa achieves remarkable performance, where the average AUC exceeds 99. Despite the impressive performance of DeBERTa in the in-domain situation, where the training data and evaluation data are from the same distribution, accessing and identifying a sufficient quantity of in-domain training data is not always possible in real-world scenarios. We further evaluate its generalization ability, where we train it on one dataset and evaluate it on another, with results shown in Figure 2. When evaluated on a dataset different from the training datasets, its performance illustrates a drop, showing poor generalization. The drop is significant between the two categories of datasets, where the average performance when trained on **Generate** and evaluated on **Rephrase** is 68.35. This underscores the challenge of training a versatile and effective machine-generated text detector.

(II) Mixture of Experts could slightly mitigate the evidence pollution in some situations, but it might harm the general performance. Table 5 illustrates a brief performance of the mixture of experts, and we present the complete results in Tables 10, 11, 12, 13 in Appendix F. Among the ten datasets, MoE could improve the performance on most datasets for different pollution strategies. Meanwhile, It works best for **Generate**, with an average of 4.18 datasets showing improvement, while **Rephrase** has an average of 2.44 datasets showing improvement. However, considering the overall performance, most of the average performance drops with the highest decline of 2.9, indicating that it cannot be adapted to various malicious text detection tasks. Meanwhile, multiple experts necessitate additional resources, where the cost per detection escalates linearly with the number of experts used, limiting this strategy in real-world scenarios.

(III) Parameter updating is the most effective defense strategy, however, the need for annotated data and the unknown when the training ends limit its practical application. Figure 3 illustrates partial important results of parameter updating with re-training data increasing, and we present the complete results in Figures 10, 11, 12, 13, and 14 in Appendix F. Besides GET and Reverse, the parameter updating strategy could significantly improve the detection performance. For example, BERT improves 1.9% on Reverse and 1.7% on Support, while DeBERTa improves 1.3% on Reverse. It is noticeable that the improvement above is the average of relative value shown in Table 3. For the original f1-score, DEFEND achieves 10.3% improvement on **Reddit** with **Echo** pollution, GET achieves 2.5% on **Politifact** with **Repeat**

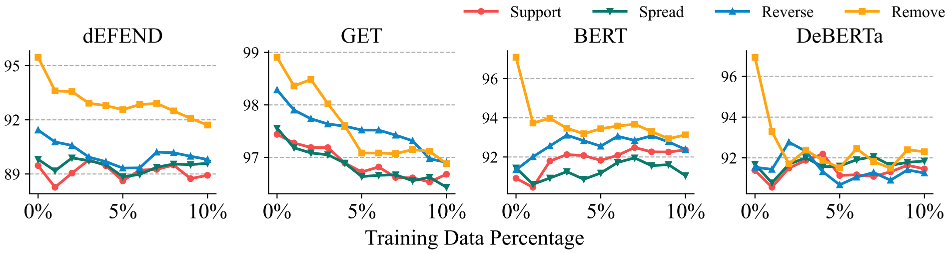

Figure 3: The performance trend of **Parameter Updating** strategy with re-training data increasing. In some situations, this strategy could significantly improve the detection performance. However, it might fail when meets **Basic** pollution such as Reverse or models that are already trained well such as GET. Meanwhile, the need for annotated data and the unknown when the training ends limit its practical application.

Pollution	DEFEND # of $\uparrow$	DEFEND Δ	GET # of $\uparrow$	GET Δ	BERT # of $\uparrow$	BERT Δ	DEBERTA # of $\uparrow$	DEBERTA Δ
Remove	1	2.9 $\downarrow$	0	1.4 $\downarrow$	0	1.3 $\downarrow$	0	1.6 $\downarrow$
Repeat	8	2.3 $\uparrow$	4	0.2 $\downarrow$	-	-	-	-
Rephrase	3	2.4 $\downarrow$	1	0.9 $\downarrow$	3	0.3 $\downarrow$	4	0.4 $\downarrow$
Rewrite	3	1.8 $\downarrow$	2	0.5 $\downarrow$	3	0.6 $\downarrow$	1	1.3 $\downarrow$
Reverse	2	0.5 $\downarrow$	3	0.4 $\downarrow$	3	0.4 $\downarrow$	1	0.6 $\downarrow$
Modify	3	1.5 $\downarrow$	1	0.7 $\downarrow$	4	0.3 $\downarrow$	2	0.6 $\downarrow$
Vanilla	3	0.2 $\uparrow$	2	0.2 $\downarrow$	4	0.1 $\downarrow$	3	0.1 $\downarrow$
Support	4	0.9 $\uparrow$	4	0.0 $\downarrow$	5	0.1 $\downarrow$	6	0.2 $\uparrow$
Oppose	6	0.4 $\uparrow$	5	0.2 $\downarrow$	3	0.2 $\downarrow$	3	0.1 $\downarrow$
Publisher	7	1.8 $\uparrow$	5	0.4 $\downarrow$	3	0.1 $\downarrow$	4	0.1 $\uparrow$
Echo	4	0.2 $\downarrow$	2	0.5 $\downarrow$	5	0.2 $\downarrow$	5	0.0 $\uparrow$
Makeup	6	1.6 $\uparrow$	5	0.4 $\downarrow$	1	0.3 $\downarrow$	6	0.1 $\downarrow$
Amplify	5	0.3 $\uparrow$	3	0.4 $\downarrow$	5	0.0 $\downarrow$	3	0.1 $\downarrow$

Table 5: The performance of **Mixture of Experts**. For short, "# of $\uparrow$" denotes the number of datasets that improve performance out of 10, and "Δ" denotes the changes of average relative values shown in Table 3, and "-" denotes that this strategy is not suitable for this model. This strategy could slightly improve the performance in some datasets, but the general improvement is not obvious and may even harm the detection ability.

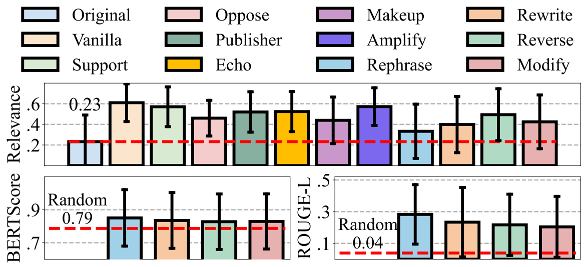

Figure 4: Evaluation of the manipulated evidence. We evaluate the relevance between social text and corresponding evidence and the semantic-level and word-level similarity between original and rephrased evidence. The polluted evidence is of high quality.

pollution, BERT achieves 17.9% on **Twitter16** with **Publisher** pollution, and DEBERTA achieves 36.9% on **Twitter16** with **Publisher** pollution, as shown in Appendix F. Although this strategy could significantly improve performance, it needs more annotated data or professional feedback to re-train the parameter, about 6-7% of the initial training data. Meanwhile, it is difficult to determine when to start or stop updating parameters since there is no more data to verify the performance. These two limitations restrict the development of this strategy to online malicious social text detection, which requires fast updating and responses.

6 Analysis

(I) The manipulated evidence is of high quality. We employ SimCSE (Gao et al., 2021) to evaluate the relevance between social text and corresponding evidence and employ BERTscore (Zhang et al., 2020) and ROUGE-L (Lin, 2004) to evaluate the semantic-level and word-level similarity between original and rephrased evidence. Figure 4 illustrates that the relevance of polluted evidence even exceeds the original. The **Generate** with an average value of 0.528 is higher than the **Rephrase** with an average value of 0.412. We speculate that LLMs could follow instructions to generate related evidence while humans tend to express their opinions unrelated to the original text. Meanwhile, the rephrased evidence is similar to the original in both semantic and word levels, with higher similarities than the randomly selected evidence pairs. We further conduct a human evaluation to check which types of evidence are of high quality. The results show that 12 out of 29 prefer generated evidence to original and 14 out of 29 prefer rephrased evidence to original. We speculate that online social users struggle to distinguish manipulated and original evidence, especially the rephrase type.

(II) Evidence pollution harms model calibration thus declining prediction trustworthiness. Robust detectors should provide a prediction and a well-calibrated confidence score to facilitate content moderation. We evaluate how well detectors are calibrated with original and polluted evidence using Expected Calibration Error (ECE) (Guo et al., 2017). Figure 5 illustrates partial results, and we

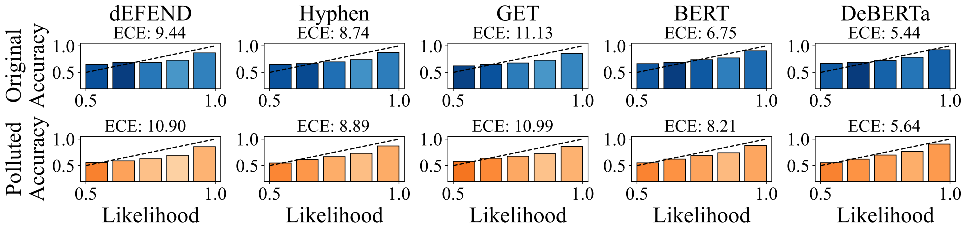

Figure 5: Calibration of existing detectors with the original and polluted evidence. ECE denotes expected calibration error, the lower the better. The dashed line indicates perfect calibration, while the color of the bar is darker when it is closer to perfect calibration. Evidence pollution could harm the model calibration.

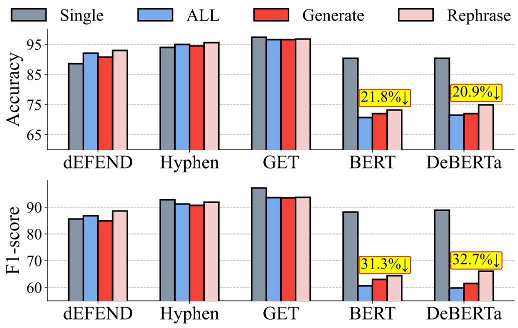

Figure 6: Performance of detectors when the pollution strategies collaborate. For short, "Single" denotes the best pollution strategies for a specific detector, "ALL" denotes the ensemble of all LLM-based strategies, and "Generate" and "Rephrase" denote the ensemble of corresponding strategies. The ensemble of evidence pollution would amplify the negative impact.

present more results in Figure 15 in Appendix G. It is demonstrated that polluted evidence harms calibration and increases ECE by up to 21.6%, while encoder-based LMs are the most well-calibrated.

(III) The ensemble of evidence pollution would amplify the negative impact. Figure 6 illustrates the performance of detectors when the pollution strategies collaborate. Encoder-based LMs are more sensitive to the ensemble, where BERT drops up to 21.8% and DEBERTA drops up to 20.9% for accuracy. Other detectors are more robust but also suffer from slight performance drops.

7 Related Work

Identifying malicious social text is critical for ensuring online safety. Researchers work on detecting fake news (Yue et al., 2023; Mendes et al., 2023; Ma et al., 2024b), identifying rumors (Kim et al., 2023; Yang et al., 2024), countering hate speech (Singh and Thakur, 2024; Tonneau et al., 2024; Lee et al., 2024), and recognizing sarcasm (Min et al., 2023; Chen et al., 2024b). Intuitive works employ technologies such as augmentation (Kim et al., 2024; Lee et al., 2024), recurrent neural networks (Shu et al., 2019), and transformer (Tian et al., 2023; Nguyen, 2024) enhanced with emotion (Zhang et al., 2021), opinions (Zong et al., 2024), semantics (Ahn et al., 2024), and logical rules (Clarke et al., 2023; Chen et al., 2023) to analyze social text content. To counter disguised content, evidence-enhanced models are proposed, utilizing external knowledge such as similar content (Sheng et al., 2022; Qi et al., 2023), comments (Yu et al., 2023; Yang et al., 2023), user (Shu et al., 2018; Dou et al., 2021), and multiply modalities (Cao et al., 2020; Tiwari et al., 2023) and then employing networks like graph neural networks (Ghosh et al., 2023; Jing et al., 2023) to fuse them.

Aside from remarkable abilities to standard NLP tasks, LLMs show great potential to conduct content moderation, such as countering social bot detection (Feng et al., 2024), misinformation (Russo et al., 2023; Yue et al., 2024; Ma et al., 2024a; Liu et al., 2024a; Su et al., 2024), hate speech (Nguyen et al., 2023; Yadav et al., 2024; Zheng et al., 2024). However, LLMs' misuse introduces risks of malicious text generation (Pelrine et al., 2023; Huang et al., 2023; Chen and Shu, 2024; Wu et al., 2024). Existing research explores the influence of misinformation (Pan et al., 2023; Goldstein et al., 2023; Xu et al., 2024) and how to detect machine-generated text (Mitchell et al., 2023). We explored the risks of evidence pollution in malicious social text detection and potential defense strategies, bridging the gap between existing works.

8 Conclusion

We explore LLMs' potential evidence pollution risks, which confuse evidence-enhanced malicious social text detectors. We design three types of manipulation strategies including thirteen methods and propose three defense strategies from both the data and model sides. Extensive experiments illustrate that evidence pollution poses a profound threat, which remains challenging to fully mitigate by employing existing defense strategies.

Limitation

While our proposed pollution strategies and defense strategies are generic, we focus on the comments, which are the most widely used. We believe the extensive experiment results on ten datasets across four malicious social text detection tasks could demonstrate our key contributions.

More recent works might employ the evidence graphs, such as the comments on other comments or user following graphs, to enhance detection performance. This paper focuses on the comments directly on the social text and the textual information instead of graph information. We also believe the extensive experiments of seven strong detectors could demonstrate our key contributions.

We expect to explore the risks of LLMs in manipulating other types of evidence and graph structure, as well as the corresponding defense strategies.

Ethics Statement

Identifying malicious social text on social platforms ensures online safety. This paper aims to explore the risks of LLMs in manipulating evidence to compromise evidence-enhanced detectors and develop potential defense strategies to mitigate evidence pollution, while also increasing the risks of dual use. We aim to mitigate such dual use by employing controlled access to our research data, making sure that the data is only employed for research purposes. Meanwhile, our research reveals the vulnerability of existing detectors to evidence pollution. Thus we argue that the decision of the existing detectors should be considered as an initial screen of malicious content, while content moderation decisions should be made with related experts.

We argue that before employing evidence to enhance malicious social text detection, fact-checking is needed to ensure the credibility of the evidence. Meanwhile, to increase the reliability of evidence-enhance detectors, increasing the explainability, such as giving out which evidence leads to the predictions, is critical.

We mainly employ LLMs to rewrite existing evidence or generate fabricated evidence with predetermined malicious intent to compromise detectors. We do not directly employ LLMs to generate malicious content, and we also argue that LLMs should not be employed to generate malicious content, where researchers should make an effort to limit it. Meanwhile, due to the inherent social bias and hallucinations of LLMs, the polluted evidence inevitably contains biased content, such as hate speech or misinformation. We emphasize that the data can only be used for research purposes.

References

Hyeseon Ahn, Youngwook Kim, Jungin Kim, and Yo-Sub Han. 2024. SharedCon: Implicit hate speech detection using shared semantics. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 10444–10455, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Arnav Arora, Preslav Nakov, Momchil Hardalov, Sheikh Muhammad Sarwar, Vibha Nayak, Yoan Dinkov, Dimitrina Zlatkova, Kyle Dent, Ameya Bhatawdekar, Guillaume Bouchard, et al. 2023. Detecting harmful content on online platforms: what platforms need vs. where research efforts go. *ACM Computing Surveys*, 56(3):1–17.

Guangsheng Bao, Yanbin Zhao, Zhiyang Teng, Linyi Yang, and Yue Zhang. 2024. Fast-detectgpt: Efficient zero-shot detection of machine-generated text via conditional probability curvature. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

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

Juan Cao, Peng Qi, Qiang Sheng, Tianyun Yang, Junbo Guo, and Jintao Li. 2020. Exploring the role of visual content in fake news detection. *Disinformation, Misinformation, and Fake News in Social Media: Emerging Research Challenges and Opportunities*, pages 141–161.

Canyu Chen and Kai Shu. 2024. Can llm-generated misinformation be detected? In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Ting-Chih Chen, Chia-Wei Tang, and Chris Thomas. 2024a. Metasumperceiver: Multimodal multi-document evidence summarization for fact-checking. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8742–8757.

Ziwei Chen, Linmei Hu, Weixin Li, Yingxia Shao, and Liqiang Nie. 2023. Causal intervention and counterfactual reasoning for multi-modal fake news detection. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 627–638.

Zixin Chen, Hongzhan Lin, Ziyang Luo, Mingfei Cheng, Jing Ma, and Guang Chen. 2024b. CofiPara: A coarse-to-fine paradigm for multimodal sarcasm target identification with large multimodal models. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 9663–9687, Bangkok, Thailand. Association for Computational Linguistics.

Christopher Clarke, Matthew Hall, Gaurav Mittal, Ye Yu, Sandra Sajeev, Jason Mars, and Mei Chen. 2023. Rule by example: Harnessing logical rules for explainable hate speech detection. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 364–376.

Leon Derczynski, Kalina Bontcheva, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Arkaitz Zubiaga. 2017. Semeval-2017 task 8: Rumoureval: Determining rumour veracity and support for rumours. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 69–76.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Chenhe Dong, Yinghui Li, Haifan Gong, Miaoxin Chen, Junxin Li, Ying Shen, and Min Yang. 2022. A survey of natural language generation. *ACM Computing Surveys*, 55(8):1–38.

Yingtong Dou, Kai Shu, Congying Xia, Philip S Yu, and Lichao Sun. 2021. User preference-aware fake news detection. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, pages 2051–2055.

Shangbin Feng, Herun Wan, Ningnan Wang, Zhaoxuan Tan, Minnan Luo, and Yulia Tsvetkov. 2024. What does the bot say? opportunities and risks of large language models in social media bot detection. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3580–3601, Bangkok, Thailand. Association for Computational Linguistics.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910.

Debanjan Ghosh, Avijit Vajpayee, and Smaranda Muresan. 2020. A report on the 2020 sarcasm detection shared task. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 1–11.

Shreya Ghosh and Prasenjit Mitra. 2023. Catching lies in the act: A framework for early misinformation detection on social media. In *Proceedings of the 34th ACM Conference on Hypertext and Social Media*, pages 1–12.

Sreyan Ghosh, Manan Suri, Purva Chiniya, Utkarsh Tyagi, Sonal Kumar, and Dinesh Manocha. 2023. Cosyn: Detecting implicit hate speech in online conversations using a context synergized hyperbolic network. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6159–6173.

Josh A Goldstein, Girish Sastry, Micah Musser, Renee DiResta, Matthew Gentzel, and Katerina Sedova. 2023. Generative language models and automated influence operations: Emerging threats and potential mitigations. *arXiv preprint arXiv:2301.04246*.

Karish Grover, SM Angara, Md Shad Akhtar, and Tanmoy Chakraborty. 2022. Public wisdom matters! discourse-aware hyperbolic fourier co-attention for social text classification. *Advances in Neural Information Processing Systems*, 35:9417–9431.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR.

Quanjiang Guo, Zhao Kang, Ling Tian, and Zhouguo Chen. 2023. Tiefake: Title-text similarity and emotion-aware fake news detection. In *2023 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7. IEEE.

Momchil Hardalov, Arnav Arora, Preslav Nakov, and Isabelle Augenstein. 2022. A survey on stance detection for mis-and disinformation identification. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1259–1277.

Philipp Hartl and Udo Kruschwitz. 2022. Applying automatic text summarization for fake news detection. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2702–2713.

Kadhim Hayawi, Sakib Shahriar, Mohamed Adel Serhani, Ikbal Taleb, and Sujith Samuel Mathew. 2022. Anti-vax: a novel twitter dataset for covid-19 vaccine misinformation detection. *Public health*, 203:23–30.

Bing He, Mustaque Ahamad, and Srijan Kumar. 2023a. Reinforcement learning-based counter-misinformation response generation: a case study of covid-19 vaccine misinformation. In *Proceedings of the ACM Web Conference 2023*, pages 2698–2709.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023b. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Maryam Heidari, Samira Zad, Parisa Hajibabaee, Masoud Malekzadeh, SeyyedPooya HekmatiAthar, Ozlem Uzuner, and James H Jones. 2021. Bert model for fake news detection based on social bot activities in the covid-19 pandemic. In *2021 IEEE 12th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)*, pages 0103–0109. IEEE.

Xuming Hu, Zhijiang Guo, Junzhe Chen, Lijie Wen, and Philip S Yu. 2023. Mr2: A benchmark for multimodal retrieval-augmented rumor detection in social media. In *Proceedings of the 46th international ACM SIGIR conference on research and development in information retrieval*, pages 2901–2912.

Kung-Hsiang Huang, Kathleen Mckeown, Preslav Nakov, Yejin Choi, and Heng Ji. 2023. Faking fake news for real fake news detection: Propaganda-loaded training data generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14571–14589.

Álvaro Huertas-García, Alejandro Martín, Javier Huertas-Tato, and David Camacho. 2023. Countering malicious content moderation evasion in online social networks: Simulation and detection of word camouflage. *Applied Soft Computing*, 145:110552.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Liqiang Jing, Xuemeng Song, Kun Ouyang, Mengzhao Jia, and Liqiang Nie. 2023. Multi-source semantic graph-based multimodal sarcasm explanation generation. In *The 61st Annual Meeting Of The Association For Computational Linguistics*.

Anna-Katharina Jung, Björn Ross, and Stefan Stieglitz. 2020. Caution: Rumors ahead—a case study on the debunking of false information on twitter. *Big Data & Society*, 7(2):2053951720980127.

Jaehoon Kim, Seungwan Jin, Sohyun Park, Someen Park, and Kyungsik Han. 2024. Label-aware hard negative sampling strategies with momentum contrastive learning for implicit hate speech detection. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 16177–16188, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Jongin Kim, Byeo Rhee Bak, Aditya Agrawal, Jiaxi Wu, Veronika Wirtz, Traci Hong, and Derry Wijaya. 2023. Covid-19 vaccine misinformation in middle income countries. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3903–3915.

Loukas Konstantinou and Evangelos Karapanos. 2023. Nudging for online misinformation: a design inquiry. In *Companion Publication of the 2023 Conference on Computer Supported Cooperative Work and Social Computing*, pages 69–75.

Nayeon Lee, Chani Jung, Junho Myung, Jiho Jin, Jose Camacho-Collados, Juho Kim, and Alice Oh. 2024. Exploring cross-cultural differences in english hate speech annotations: From dataset construction to analysis. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4205–4224.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Hongzhan Lin, Zixin Chen, Ziyang Luo, Mingfei Cheng, Jing Ma, and Guang Chen. 2024. Cofipara: A coarse-to-fine paradigm for multimodal sarcasm target identification with large multimodal models. *arXiv preprint arXiv:2405.00390*.

Hui Liu, Wenya Wang, Haoru Li, and Haoliang Li. 2024a. Teller: A trustworthy framework for explainable, generalizable and controllable fake news detection. *arXiv preprint arXiv:2402.07776*.

Tianrui Liu, Qi Cai, Changxin Xu, Bo Hong, Fanghao Ni, Yuxin Qiao, and Tsungwei Yang. 2024b. Rumor detection with a novel graph neural network approach. *Academic Journal of Science and Technology*, 10(1):305–310.

Jason Lucas, Adaku Uchendu, Michiharu Yamashita, Jooyoung Lee, Shaurya Rohatgi, and Dongwon Lee. 2023. Fighting fire with fire: The dual role of llms in crafting and detecting elusive disinformation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14279–14305.

Jing Ma, Wei Gao, and Kam-Fai Wong. 2018. Rumor detection on twitter with tree-structured recursive neural networks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1980–1989.

Weicheng Ma, Chunyuan Deng, Aram Moossavi, Lili Wang, Soroush Vosoughi, and Diyi Yang. 2024a. Simulated misinformation susceptibility (smists): Enhancing misinformation research with large language model simulations. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 2774–2788.

Zihan Ma, Minnan Luo, Hao Guo, Zhi Zeng, Yiran Hao, and Xiang Zhao. 2024b. Event-radar: Event-driven multi-view learning for multimodal fake news detection. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5809–5821.

Alex Troy Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *The 61st Annual Meeting Of The Association For Computational Linguistics*.

Thomas Mandl, Sandip Modha, Prasenjit Majumder, Daksh Patel, Mohana Dave, Chintak Mandlia, and Aditya Patel. 2019. Overview of the hasoc track at fire 2019: Hate speech and offensive content identification in indo-european languages. In *Proceedings of the 11th annual meeting of the Forum for Information Retrieval Evaluation*, pages 14–17.

Ethan Mendes, Yang Chen, Wei Xu, and Alan Ritter. 2023. Human-in-the-loop evaluation for early misinformation detection: A case study of covid-19 treatments. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15817–15835.

Changrong Min, Ximing Li, Liang Yang, Zhilin Wang, Bo Xu, and Hongfei Lin. 2023. Just like a human would, direct access to sarcasm augmented with potential result and reaction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10172–10183.

Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D Manning, and Chelsea Finn. 2023. Detectgpt: Zero-shot machine-generated text detection using probability curvature. In *International Conference on Machine Learning*, pages 24950–24962. PMLR.

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. *arXiv preprint arXiv:2405.16631*.

Luan Thanh Nguyen. 2024. Vihatet5: Enhancing hate speech detection in vietnamese with a unified text-to-text transformer model. *arXiv preprint arXiv:2405.14141*.

Tin Nguyen, Jiannan Xu, Aayushi Roy, Hal Daumé III, and Marine Carpuat. 2023. Towards conceptualization of "fair explanation": Disparate impacts of anti-asian hate speech explanations on content moderators. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9696–9717.

Tuc Nguyen and Thai Le. 2024. Generalizability of mixture of domain-specific adapters from the lens of signed weight directions and its application to effective model pruning. *arXiv preprint arXiv:2402.10639*.

Yikang Pan, Liangming Pan, Wenhu Chen, Preslav Nakov, Min-Yen Kan, and William Wang. 2023. On the risk of misinformation pollution with large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1389–1403.

Kellin Pelrine, Anne Imouza, Camille Thibault, Meilina Reksoprodjo, Caleb Gupta, Joel Christoph, Jean-François Godbout, and Reihaneh Rabbany. 2023. Towards reliable misinformation mitigation: Generalization, uncertainty, and gpt-4. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6399–6429.

Gordon Pennycook, Tyrone D Cannon, and David G Rand. 2018. Prior exposure increases perceived accuracy of fake news. *Journal of experimental psychology: general*, 147(12):1865.

Kashyap Popat, Subhabrata Mukherjee, Andrew Yates, and Gerhard Weikum. 2018. Declare: Debunking fake news and false claims using evidence-aware deep learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 22–32.

Peng Qi, Yuyang Zhao, Yufeng Shen, Wei Ji, Juan Cao, and Tat-Seng Chua. 2023. Two heads are better than one: Improving fake news video detection by correlating with neighbors. In *The 61st Annual Meeting Of The Association For Computational Linguistics*.

Daniel Russo, Shane Kaszefski-Yaschuk, Jacopo Staiano, and Marco Guerini. 2023. Countering misinformation via emotional response generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11476–11492.

Omar Shaikh, Hongxin Zhang, William Held, Michael Bernstein, and Diyi Yang. 2023. On second thought, let's not think step by step! bias and toxicity in zero-shot reasoning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4454–4470.

Lanyu Shang, Yang Zhang, Zhenrui Yue, YeonJung Choi, Huimin Zeng, and Dong Wang. 2024. A domain adaptive graph learning framework to early detection of emergent healthcare misinformation on social media. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 18, pages 1408–1421.

Qiang Sheng, Juan Cao, Xueyao Zhang, Rundong Li, Danding Wang, and Yongchun Zhu. 2022. Zoom out and observe: News environment perception for fake news detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4543–4556.

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.

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.

Kai Shu, Suhang Wang, and Huan Liu. 2018. Understanding user profiles on social media for fake news detection. In *2018 IEEE conference on multimedia information processing and retrieval (MIPR)*, pages 430–435. IEEE.

Akshay Singh and Rahul Thakur. 2024. Generalizable multilingual hate speech detection on low resource indian languages using fair selection in federated learning. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7204–7214.

Jinyan Su, Claire Cardie, and Preslav Nakov. 2024. Adapting fake news detection to the era of large language models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 1473–1490.

Yuan Tian, Nan Xu, Ruike Zhang, and Wenji Mao. 2023. Dynamic routing transformer network for multimodal sarcasm detection. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2468–2480.

Yuanhe Tian, Fei Xia, and Yan Song. 2024. Dialogue summarization with mixture of experts based on large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7143–7155.

Divyank Tiwari, Diptesh Kanojia, Anupama Ray, Apoorva Nunna, and Pushpak Bhattacharyya. 2023. Predict and use: Harnessing predicted gaze to improve multimodal sarcasm detection. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15933–15948.

Manuel Tonneau, Pedro Quinta De Castro, Karim Lasri, Ibrahim Farouq, Lakshmi Subramanian, Victor Orozco-Olvera, and Samuel Fraiberger. 2024. NaijaHate: Evaluating hate speech detection on Nigerian Twitter using representative data. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9020–9040, Bangkok, Thailand. Association for Computational Linguistics.

Adaku Uchendu, Jooyoung Lee, Hua Shen, Thai Le, Dongwon Lee, et al. 2023. Does human collaboration enhance the accuracy of identifying llm-generated deepfake texts? In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 11, pages 163–174.

Vivek Verma, Eve Fleisig, Nicholas Tomlin, and Dan Klein. 2024. Ghostbuster: Detecting text ghostwritten by large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1702–1717.

Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *science*, 359(6380):1146–1151.

Herun Wan, Shangbin Feng, Zhaoxuan Tan, Heng Wang, Yulia Tsvetkov, and Minnan Luo. 2024. DELL: Generating reactions and explanations for LLM-based misinformation detection. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 2637–2667, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Xinyu Wang, Jiayi Li, and Sarah Rajtmajer. 2024a. Inside the echo chamber: Linguistic underpinnings of misinformation on twitter. In *Proceedings of the 16th ACM Web Science Conference*, pages 31–41.

Yichen Wang, Shangbin Feng, Abe Bohan Hou, Xiao Pu, Chao Shen, Xiaoming Liu, Yulia Tsvetkov, and Tianxing He. 2024b. Stumbling blocks: Stress testing the robustness of machine-generated text detectors under attacks. *arXiv preprint arXiv:2402.11638*.

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.

Rongwu Xu, Brian Lin, Shujian Yang, Tianqi Zhang, Weiyan Shi, Tianwei Zhang, Zhixuan Fang, Wei Xu, and Han Qiu. 2024. The earth is flat because...: Investigating LLMs' belief towards misinformation via persuasive conversation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16259–16303, Bangkok, Thailand. Association for Computational Linguistics.

Weizhi Xu, Junfei Wu, Qiang Liu, Shu Wu, and Liang Wang. 2022. Evidence-aware fake news detection with graph neural networks. In *Proceedings of the ACM web conference 2022*, pages 2501–2510.

Neemesh Yadav, Sarah Masud, Vikram Goyal, Md Shad Akhtar, and Tanmoy Chakraborty. 2024. Tox-BART: Leveraging toxicity attributes for explanation generation of implicit hate speech. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 13967–13983, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Ruichao Yang, Wei Gao, Jing Ma, Hongzhan Lin, and Bo Wang. 2024. Reinforcement tuning for detecting stances and debunking rumors jointly with large language models. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 13423–13439, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Ruichao Yang, Wei Gao, Jing Ma, Hongzhan Lin, and Zhiwei Yang. 2023. Wsdms: Debunk fake news via weakly supervised detection of misinforming sentences with contextualized social wisdom. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1525–1538.

Xinchen Yu, Ashley Zhao, Eduardo Blanco, and Lingzi Hong. 2023. A fine-grained taxonomy of replies to hate speech. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7275–7289.

Xin Yuan, Jie Guo, Weidong Qiu, Zheng Huang, and Shujun Li. 2023. Support or refute: Analyzing the stance of evidence to detect out-of-context mis-and disinformation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4268–4280.

Zhenrui Yue, Huimin Zeng, Yimeng Lu, Lanyu Shang, Yang Zhang, and Dong Wang. 2024. Evidence-driven retrieval augmented response generation for online misinformation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5628–5643.

Zhenrui Yue, Huimin Zeng, Yang Zhang, Lanyu Shang, and Dong Wang. 2023. Metaadapt: Domain adaptive few-shot misinformation detection via meta learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5223–5239.

Min Zhang, Jianfeng He, Taoran Ji, and Chang-Tien Lu. 2024. Don't go to extremes: Revealing the excessive sensitivity and calibration limitations of llms in implicit hate speech detection. *arXiv preprint arXiv:2402.11406*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

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

Hao Zhao, Zihan Qiu, Huijia Wu, Zili Wang, Zhaofeng He, and Jie Fu. 2024. Hypermoe: Towards better mixture of experts via transferring among experts. *arXiv preprint arXiv:2402.12656*.

Jiangrui Zheng, Xueqing Liu, Mirazul Haque, Xing Qian, Guanqun Yang, and Wei Yang. 2024. Hatemoderate: Testing hate speech detectors against content moderation policies. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2691–2710.

Jonathan Zheng, Ashutosh Baheti, Tarek Naous, Wei Xu, and Alan Ritter. 2022. Stanceosaurus: Classifying stance towards multicultural misinformation. In *Proceedings of the 2022 conference on empirical methods in natural language processing*, pages 2132–2151.

Ming Zhou, Dan Zhang, Yuandong Wang, Yangli-ao Geng, Yuxiao Dong, and Jie Tang. 2024. Lgb: Language model and graph neural network-driven social bot detection. *arXiv preprint arXiv:2406.08762*.

Linlin Zong, Jiahui Zhou, Wenmin Lin, Xinyue Liu, Xianchao Zhang, and Bo Xu. 2024. Unveiling opinion evolution via prompting and diffusion for short video fake news detection. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 10817–10826.

A Case Study

We present case studies of each evidence pollution strategy in Tables 14 and 15. Note that these cases are all malicious social texts from the public dataset **Politifact**, and we have concealed personal private information and hate speech as much as possible. We could summarize the characteristics of each evidence pollution strategy as follows:

- **Remove** simply removes some related evidence, where the removed evidence might provide useful signals to identify the malicious content. It is straightforward but difficult to implement in practice due to platform rules.
- **Repeat** aims to repeat unified evidence to amplify its influence. It is easily detected by the platforms through the text-matching algorithm.
- **Rephrase** rephrases existing evidence without any additional intents. It is just like a baseline for **Rephrase Evidence**.
- **Rewrite** rewrites existing evidence intending to make the corresponding social text like a normal one. Thus, LLMs might generate some clarifications in the evidence.
- **Reverse** reverses the stance in existing evidence, thus it might completely replace the content related to the stance.
- **Modify** adds fabricated facts to make the social text human-like.
- **Vanilla** simply generates related evidence of the corresponding social text. It is just like a baseline for **Generate Evidence**.
- **Support** generates evidence with the predetermined support stance.
- **Oppose** generates evidence with the predetermined opposing stance.

- **Publisher** simulates the social text publishers to post comments to promote the original social text. For example, LLMs could generate some hashtags.
- **Echo** aims to create echo chambers, where it would post comments with similar semantics. It might be more difficult to be detected by the platforms.
- **Makeup** generates evidence intending to make the corresponding social text like a normal one.
- **Amplify** aims to generate evidence to promote the spread of corresponding social text. Thus LLMs might generate hashtags and employ interrogative sentences.

These cases show that the polluted evidence is of high quality, where LLMs could follow the instructions to rewrite or generate highly relevant evidence, confusing existing evidence-enhanced malicious social text detectors.

B Metric Set

We mainly employ accuracy, macro f1-score, $\mathrm{AR}_{\mathrm{acc}}$ and $\mathrm{AR}_{\mathrm{F1}}$, and AUC as metrics. We introduce each of the metrics and the reasons to employ them:

- Accuracy and macro f1-score are widely used metrics for classification tasks. Thus we employ them to evaluate the general performance of detectors. For the accuracy, we employ it in Tables 2, 10, 11, 12, and 13, and in Figures 3, 6, 7, 8, 9, 10, 11, 12, 13, and 14. For the macro f1-score, we employ it in Tables 9, 10, 11, 12, and 13, and Figure 6.
- $\mathrm{AR}_{\mathrm{acc}}$ and $\mathrm{AR}_{\mathrm{F1}}$ are proposed to evaluate the influence of pollution strategies. Given a specific detector and a pollution strategy, we assume the original performance (accuracy or macro f1-score) is $\{f_i\}_{i=1}^{N}$, where N is the number of datasets (we employ 10 datasets), and the performance after pollution is $\{\tilde{f}_i\}_{i=1}^{N}$. The AR is calculated as:

$$\mathrm{AR} = \frac{1}{N}\sum_{i=1}^{N}\frac{\tilde{f}_i}{f_i}.$$

  The lower the value, the more effective the pollution strategy is. Meanwhile, given an AR score, it is convenient to calculate the relative performance drop rate: $1 - \mathrm{AR}$. We employ AR in Tables 3 and 5.
- AUC is widely used in machine-generated text detection, thus we employ it to evaluate the performance of machine-generated text detectors, as well as the f1-score. We employ them in Table 4 and Figure 2.

C Baselines

We evaluate our proposed evidence pollution and defense strategies on three distinctive types of competitive detectors. The first category is **existing strong detector**, which presents the most advanced technologies, and we employ:

- dEFEND (Shu et al., 2019) conducts explainable detection by the attention weights between social text sentences and related evidence. We set the max sentence count of the social text as 8 and the max token count of each sentence as 128. We further set the max evidence count as 10 and the max token count of evidence as 128.
- HYPHEN (Grover et al., 2022) is a discourse-aware hyperbolic spectral co-attention network. It employs a novel Fourier co-attention mechanism to enhance hyperbolic graph representations, obtaining joint representations of social text and evidence. We set the max evidence to count as 10 and the max token count of social text sentence as 128. We further set the max social text sentence count as the 80th percentile for each dataset.
- GET (Xu et al., 2022) models social text and evidence as networks and captures the long-distance semantic dependency among dispersed relevant snippets via neighborhood propagation. For both social text and evidence graphs, we set the max word length as 3840 and set the window size as 5.

The second category is **encoder-based LM**, where we employ encoder-based LMs to encode social text and evidence content and then fuse their representations to conduct classification. Specifically, given a piece of social text s and its corresponding evidence $\{c_i\}_{i=1}^{m}$, we first employ encoder-based LMs $\mathrm{enc}(\cdot)$ to obtain their representations, *i.e.*,

$$\mathbf{h}_{text} = \mathrm{enc}(s),$$

$$\mathbf{h}_{evid} = \sum_{i=1}^{m}\mathrm{enc}(c_i).$$

Methods	TASK	Prompt
Generic input prompt: *Text: s*		
F3 *VaN*	Fake News	*Analyze the given text and determine if it is real or fake news.*
	Hate Speech	*Analyze the given text and determine if it is hate speech or not.*
	Rumor	*Analyze the given text and determine if it is a rumor or not a rumor.*
	Sarcasm	*Analyze the given text and determine if it is sarcasm or not.*
Generic input prompt: *Text: s Comments: $i.c_i$. Analyze the given text and related comments,*		
w/ evidence	Fake News	*and determine if it is real or fake news.*
	Hate Speech	*and determine if it is hate speech or not.*
	Rumor	*and determine if it is a rumor or not a rumor.*
	Sarcasm	*and determine if it is sarcasm or not.*

Table 6: Prompts of **LLM-based detectors**, we prompt LLMs using F3 (Lucas et al., 2023) and with evidence.

Hyper	dEFEND	HYPHEN	GET	BERT	DeBERTa
Optimizer	Adam (RiemannianAdam for HYPHEN)				
Metrics	Accuracy				
Weight Decay	1e-5				
Dropout	0.5				
Hidden Dim	256				
Learning Rate	1e-4	1e-3	1e-3	1e-4	1e-4
Batch Size	32	32	32	16	16
Only for **Politifact**, **Gossipcop**, and **RumorEval**.					
Batch Size	32	32	32	16	4

Table 7: Hyperparameters of baselines required to train.

We then concatenate them to obtain the final representation:

$$\mathbf{h} = \mathbf{h}_{text} \| \mathbf{h}_{evid}.$$

Finally, given an instance and its label y, we compute the probability of y being the correct prediction as $p(y \mid \mathcal{G}) \propto \exp(\mathrm{MLP}(\mathbf{h}))$, where $\mathrm{MLP}(\cdot)$ denotes an MLP layer. We optimize models using the cross-entropy loss and predict the most plausible label as $\arg\max_y p(y \mid \mathcal{G})$. In practice, we employ two widely-used encoder-based LMs: (i) BERT (Devlin et al., 2019) and (ii) DeBERTa (He et al., 2023b). For LMs without evidence, we directly consider $\mathbf{h}_{text}$ as $\mathbf{h}$.

The third category is **LLM-based detector**, where we prompt LLMs with F3 (Lucas et al., 2023) and evidence. The detailed prompts are presented in Table 6. In practice, we employ an open-sourced LLM MISTRAL and a close-sourced LLM CHATGPT.

D Exeriment Settings

D.1 Baseline Settings

For each baseline, we conduct ten-fold cross-validation on each dataset to obtain more robust results. We set the hyperparameters the same for each fold. Meanwhile, we run each fold five times and select the checkpoint with the best performance. For each run, we stop training when the performance on the test set does not improve for five epochs. We present the hyperparameters of existing strong detectors and encoder-based LMs in Table 7. For LLM-based Detectors, we set the max new token to count as 50 and set the temperature as zero to obtain fixed predictions.

Task	Dataset	# Text	# Malicious	Average # Evidence
Fake News	**Politifact**	415	270	7.9
	Gossipcop	2,411	1,408	7.6
	AnTiVax	3,797	932	3.6
Hate Speech	**HASOC**	712	298	2.6
Rumor	**Pheme**	6,425	2,402	7.2
	Twitter15	543	276	4.5
	Twitter16	362	163	7.2
	RumorEval*	446	138	8.1
Sarcasm	**Twitter**	5000	2500	3.6
	Reddit	4400	2200	2.5

Table 8: The statistics of the datasets. $\star$ denotes that this dataset contains additional "not verified" class.

D.2 Dataset Settings

We employ four malicious social text detection tasks including 10 datasets, *i.e.*, (i) **fake news detection**: Politicalfact, Gossipcop (Shu et al., 2020), and ANTiVax (Hayawi et al., 2022); **hate speech detection**: HASOC (Mandl et al., 2019); (iii) **rumor detection**: Pheme (Buntain and Golbeck, 2017), Twitter15, Twitter16 (Ma et al., 2018), and RumorEval (Derczynski et al., 2017); (iv) **sarcasm detection**: Twitter and Reddit (Ghosh et al., 2020).

For original content and corresponding evidence, we employ the processed data from HYPHEN (Grover et al., 2022). We randomly split them into 10 folds to support a ten-fold evaluation. To adapt to each detector and ensure a fair comparison, we randomly down-sample relevant evidence for each social text instance, where each instance contains at most ten pieces of evidence. Table 8 presents statistics of the datasets.

D.3 Evidence Pollution Settings

We employ *Mistral-7B* (Jiang et al., 2023) to rephrase and generate polluted evidence. To ensure reproducibility, we set the temperature as zero. For **Rephrase** strategy, we prompt LLMs to rephrase in three ways, however, we employ the first version in practice because their performance is similar.

D.4 Defense Strategy Settings

Machine-Generated Text Detection To construct datasets for evaluating machine-generated text detectors, we sample 200 pieces of evidence from each dataset on each pollution strategy and

Method	Fake News			Hate Speech	Rumor				Sarcasm	
	Politifact	Gossipcop	ANTiVax	HASOC	Pheme	Twitter15	Twitter16	RumorEval	Twitter	Reddit
dEFEND (Shu et al., 2019)	$81.4_{\pm5.1}$	$70.7_{\pm2.4}$	$90.1_{\pm1.8}$	$68.4_{\pm4.2}$	$79.6_{\pm0.9}$	$84.4_{\pm4.2}$	$90.6_{\pm2.8}$	$57.6_{\pm3.5}$	$75.0_{\pm1.8}$	$66.2_{\pm1.3}$
HYPHEN (Grover et al., 2022)	$88.0_{\pm6.2}$	$69.1_{\pm2.5}$	$90.6_{\pm1.8}$	$67.9_{\pm7.6}$	$81.0_{\pm1.3}$	$\underline{90.3}_{\pm5.3}$	$93.1_{\pm3.2}$	$63.2_{\pm5.0}$	$75.5_{\pm2.0}$	$67.6_{\pm2.2}$
GET (Xu et al., 2022)	$93.5_{\pm4.8}$	$74.3_{\pm2.3}$	$91.3_{\pm0.7}$	$66.9_{\pm5.1}$	$84.8_{\pm1.5}$	$\mathbf{92.2}_{\pm2.5}$	$\mathbf{94.8}_{\pm3.3}$	$63.7_{\pm5.2}$	$74.1_{\pm1.5}$	$65.9_{\pm2.2}$
BERT (Devlin et al., 2019)	$94.0_{\pm2.9}$	$\underline{76.3}_{\pm1.8}$	$93.2_{\pm1.5}$	$\mathbf{71.4}_{\pm4.7}$	$\underline{85.4}_{\pm1.3}$	$88.5_{\pm3.8}$	$\underline{93.8}_{\pm4.4}$	$\mathbf{69.0}_{\pm4.9}$	$\underline{81.0}_{\pm1.4}$	$70.1_{\pm1.9}$
BERT w/o comments	$93.1_{\pm3.7}$	$75.2_{\pm2.5}$	$92.4_{\pm1.0}$	$\underline{69.0}_{\pm5.4}$	$\mathbf{86.2}_{\pm1.8}$	$90.2_{\pm3.3}$	$\underline{93.8}_{\pm4.0}$	$\underline{66.1}_{\pm6.2}$	$79.2_{\pm1.2}$	$69.7_{\pm1.8}$
DEBERTA (He et al., 2023b)	$\mathbf{96.2}_{\pm3.5}$	$\mathbf{77.3}_{\pm1.8}$	$\mathbf{94.4}_{\pm1.6}$	$64.7_{\pm3.1}$	$80.0_{\pm1.4}$	$83.4_{\pm4.2}$	$90.0_{\pm3.9}$	$62.8_{\pm6.5}$	$\mathbf{81.8}_{\pm1.4}$	$\mathbf{73.7}_{\pm2.1}$
DEBERTA w/o comments	$\underline{96.0}_{\pm3.4}$	$74.3_{\pm3.4}$	$\underline{93.9}_{\pm1.7}$	$62.2_{\pm5.4}$	$80.9_{\pm1.0}$	$83.1_{\pm4.3}$	$91.1_{\pm4.2}$	$64.9_{\pm5.8}$	$79.7_{\pm1.1}$	$\underline{72.7}_{\pm2.0}$
MISTRAL VaN (Lucas et al., 2023)	$60.7_{\pm8.5}$	$33.1_{\pm2.7}$	$52.8_{\pm2.2}$	$44.1_{\pm4.5}$	$47.1_{\pm1.7}$	$37.7_{\pm9.4}$	$34.5_{\pm5.5}$	$30.4_{\pm10.9}$	$63.0_{\pm1.7}$	$55.7_{\pm2.1}$
MISTRAL w/ comment	$53.2_{\pm10.1}$	$39.2_{\pm4.1}$	$36.6_{\pm2.9}$	$46.0_{\pm5.1}$	$50.5_{\pm1.7}$	$36.7_{\pm6.3}$	$31.1_{\pm3.1}$	$37.1_{\pm8.4}$	$59.0_{\pm2.5}$	$51.6_{\pm1.6}$
CHATGPT VaN (Lucas et al., 2023)	$49.3_{\pm7.5}$	$29.1_{\pm2.1}$	$45.0_{\pm2.0}$	$55.6_{\pm6.0}$	$27.8_{\pm1.0}$	$39.7_{\pm5.2}$	$39.3_{\pm5.6}$	$39.1_{\pm8.9}$	$40.4_{\pm2.0}$	$37.1_{\pm1.9}$
CHATGPT w/ comments	$61.7_{\pm7.3}$	$29.2_{\pm2.2}$	$59.4_{\pm3.9}$	$56.4_{\pm5.4}$	$31.1_{\pm1.2}$	$45.6_{\pm7.7}$	$38.8_{\pm7.5}$	$23.2_{\pm6.3}$	$60.2_{\pm1.8}$	$53.4_{\pm1.9}$

Table 9: Macro f1-Score of baselines on ten datasets from four malicious text-related tasks. We conduct ten-fold cross-validation and report the mean and standard deviation to obtain a more robust conclusion. **Bold** indicates the best performance and underline indicates the second best. Evidence could provide valuable signals to enhance detection, however, LLM-based models struggle to detect malicious content.

original evidence, resulting in 2,000 sentences for each set. We then consider the polluted evidence as machine-generated data and the original evidence as human-written data and mix them, obtaining 11 datasets where each dataset contains 4,000 sentences, named by the pollution strategy, such as **Rephrase** and **Support**. We finally split each dataset into the training set, valuation set, and test set by 2:1:1. For *metric-based* methods not required to train, we evaluate it on the test set. We employ roc auc and f1-score as metrics. For DeBERTa-v3, we set batch size as 24, learning rate as 1e-4, optimizer as Adam, weight decay as 1e-5, and hidden dim as 512. For FastGPT, we employ the official implementation[2] to obtain the prediction results.

For out-of-domain evaluation of DeBERTa, we keep the parameters the same and directly evaluate DeBERTa trained on a specific dataset on another.

Mixture of Expert We set k as m, namely, if a specific social text contains m pieces of evidence, then, we consider each piece of evidence as a group, obtaining m groups. Formally, given a detector f and its fixed parameters θ, social text s, and its corresponding evidence $\{c_i\}_{i=1}^{m}$, we could obtain m predictions as:

$$y_i = \arg\max_{y} p(y \mid s, \{c_i\}, f, \theta).$$

We then obtain the final prediction as:

$$y = \arg\max_{y_j}(\sum_{i=1}^{k} \mathbf{I}(y_i = y_j)).$$

[2]https://github.com/baoguangsheng/fast-detect-gpt

We evaluate this strategy on existing **strong detectors** and **encoder-based LMs** except HYPHEN. HYPHEN extracts the reference relations from multiple pieces of evidence, thus unsuitable for this strategy and would cost huge computation resources. Meanwhile, this strategy is unsuitable for **LLM-based detectors**, where it would cost huge input tokens. Given m pieces of evidence, the consumed tokens would be increased by m times.

Parameter Updating We employ 1% to 10% data from the training set to update the model parameters for each dataset, where we set the learning rate as 1e-4, batch size as 5, weight decay as 1e-5, and optimizer as Adam. To simulate the realistic situation that required a quick response, we just retrain the model using the training data only once.

D.5 Analysis Settings

Metric-based Evaluation of Polluted Evidence We first randomly sample 100 instances from each dataset to obtain a generic evaluation. We then calculate the relevant score between social text and corresponding evidence and calculate the BERTScore and ROUGE-L between rephrased and original evidence. For the "Random" category, we shuffle the initial polluted-original evidence pairs and consider it as a baseline.

For the relevant scores, we employ the hugging face implementation[3]. For BERTScore, we employ its official implementation[4] and set rescale with baseline as False, and for ROUGE-L, we employ the python packet[5].

[3]https://huggingface.co/princeton-nlp/sup-simcse-bert-base-uncased

[4]https://github.com/Tiiiger/bert_score

[5]https://pypi.org/project/rouge-score/

Human Evaluation of Polluted Evidence We recruit 99 annotators familiar with social networking platforms to judge which comment is of higher quality for a certain social text. For each annotator, we sample 15 generate-original evidence pairs, 15 rephrase-original evidence pairs, 15 generate-rephrase evidence pairs, and 5 randomly shuffled pairs as benchmark questions where the comment with higher quality is clear. We first give each annotator a brief guideline:

Thank you for attending our human evaluation. Social media users would comment on a post to express their opinions. You are asked to check which comment is of higher quality for a certain post (comment 1 or 2). Please consider factors such as relevance to the post, tone, suitability for the social platform (for the use of hashtags), etc. Please do not consider the length and grammatical errors of the comment. If you think two comments are of equal quality, please subjectively choose the one you like.

After that, if an annotator correctly identifies 3 out of 5 benchmark questions, we accept his annotations, obtaining 29 annotations.

Calibration Settings We consider the max value of the logits after the softmax operator as the confidence scores. For example, if the output is [0.8, 0.2], then the confidence score is 0.8, and if the output is [0.25, 0.75], then the confidence score is 0.75. Figure 5 presents the model calibration when the evidence pollution strategies are mixed, while Figure 15 presents the calibration of each pollution strategy.

Pollution Ensemble Settings We directly employ majority voting to obtain the ensemble predictions by multiple pollution strategies.

E More Results of Evidence Pollution

Table 9 presents the macro f1-score of baselines, where it shows a similar trend as accuracy.

Meanwhile, we present the whole accuracy of the seven baselines on ten datasets under each pollution strategy in Figures 7, 8, and 9. We only present accuracy because macro-f1 shows similar trends as accuracy shown in Tables 2 and 9. The additional results strengthen that evidence pollution significantly compromised evidence-enhanced malicious social text detection performance.

F More Results of Defense Strategies

F.1 Mixture of Experts

Tables 10, 11, 12, and 13 present the performance of **mixture of experts** of each baseline on different datasets under different pollution strategies. We highlight the values where the strategy imitates the negative impact. The results show that this defense strategy could improve the detection performance on some datasets under some strategies. However, in some cases, this strategy might harm the performance. It strength that although the mixture of experts could improve the performance, it would introduce some noise, declining the performance.

F.2 Paramter Updating

Figures 10, 11, 12, 13, and 14 illustrate the whole results, where we present the improvements and highlight the top-ten performance. The results show that this strategy is the most successful strategy, where the improvements are the most significant. On the other hand, the need for annotated data and the unknown when the training ends limit its practical application.

G More Analysis

G.1 Human Evaluation

Among the 29 acceptable annotators, 12 out of 29 prefer generated evidence to original, 14 out of 29 prefer rephrased evidence to original, and 17 out of 29 prefer rephrased evidence to generated.

G.2 Calibration

We present the calibration of each baseline under different pollution strategies in Figure 15. It illustrates that any pollution strategy could harm model calibration.

Pollution	Fake News						Hate Speech		Rumor								Sarcasm			
	Politifact		Gossipcop		ANTiVax		HASOC		Pheme		Twitter15		Twitter16		RumorEval		Twitter		Reddit	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
Vanilla	74.2$_{\pm8.3}$	70.8$_{\pm7.8}$	67.0$_{\pm4.1}$	63.4$_{\pm5.2}$	92.0$_{\pm2.0}$	89.1$_{\pm2.7}$	59.7$_{\pm7.8}$	50.0$_{\pm10.2}$	80.4$_{\pm1.2}$	78.9$_{\pm1.4}$	77.2$_{\pm7.1}$	77.0$_{\pm7.2}$	77.6$_{\pm10.5}$	76.5$_{\pm11.2}$	48.6$_{\pm6.8}$	45.8$_{\pm7.4}$	71.4$_{\pm1.9}$	71.4$_{\pm1.9}$	54.5$_{\pm2.3}$	47.2$_{\pm6.9}$
	71.5$_{\pm8.2}$	68.3$_{\pm7.3}$	67.4$_{\pm4.7}$	61.5$_{\pm7.0}$	91.5$_{\pm1.7}$	88.5$_{\pm2.4}$	61.9$_{\pm6.5}$	54.5$_{\pm8.6}$	80.0$_{\pm1.4}$	78.6$_{\pm1.5}$	76.9$_{\pm9.5}$	76.5$_{\pm9.7}$	76.5$_{\pm9.7}$	75.9$_{\pm9.5}$	46.1$_{\pm10.2}$	37.2$_{\pm11.5}$	71.3$_{\pm2.5}$	71.2$_{\pm2.5}$	59.9$_{\pm2.2}$	57.7$_{\pm4.2}$
	3.6% ↓	3.5% ↓	0.6% ↑	3.1% ↓	0.5% ↓	0.7% ↓	3.8% ↑	9.0% ↑	0.5% ↓	0.4% ↓	0.3% ↓	0.6% ↓	1.4% ↓	0.8% ↓	5.1% ↓	18.7% ↓	0.2% ↓	0.3% ↓	9.9% ↑	22.2% ↑
Publisher	70.6$_{\pm9.0}$	67.8$_{\pm7.8}$	67.5$_{\pm4.8}$	64.1$_{\pm5.5}$	91.7$_{\pm1.6}$	88.8$_{\pm2.2}$	58.8$_{\pm5.9}$	47.0$_{\pm7.9}$	80.0$_{\pm1.3}$	78.5$_{\pm1.5}$	73.8$_{\pm8.7}$	73.7$_{\pm8.7}$	77.0$_{\pm11.5}$	76.1$_{\pm11.8}$	48.4$_{\pm8.6}$	45.6$_{\pm8.7}$	71.0$_{\pm2.6}$	71.0$_{\pm2.7}$	54.5$_{\pm3.6}$	47.2$_{\pm7.9}$
	69.9$_{\pm9.1}$	65.8$_{\pm9.6}$	67.2$_{\pm5.0}$	60.9$_{\pm7.6}$	91.8$_{\pm1.6}$	88.8$_{\pm2.2}$	62.5$_{\pm7.2}$	54.5$_{\pm8.2}$	80.1$_{\pm1.3}$	78.7$_{\pm1.5}$	79.0$_{\pm8.2}$	78.6$_{\pm8.5}$	77.3$_{\pm8.4}$	76.8$_{\pm8.0}$	47.7$_{\pm10.2}$	39.2$_{\pm12.5}$	71.4$_{\pm2.5}$	71.3$_{\pm2.6}$	60.2$_{\pm2.1}$	58.0$_{\pm4.2}$
	1.0% ↓	2.9% ↓	0.6% ↓	4.9% ↓	0.1% ↑	0.0% ↑	6.2% ↑	15.9% ↑	0.2% ↑	0.3% ↑	7.0% ↑	6.6% ↑	0.4% ↑	0.9% ↑	1.5% ↓	13.9% ↓	0.5% ↑	0.4% ↑	10.4% ↑	22.9% ↑
Echo	75.2$_{\pm11.0}$	72.2$_{\pm11.1}$	68.1$_{\pm4.0}$	64.3$_{\pm5.3}$	91.8$_{\pm2.4}$	88.8$_{\pm3.2}$	59.0$_{\pm7.8}$	49.7$_{\pm8.9}$	80.5$_{\pm0.7}$	78.9$_{\pm0.9}$	78.4$_{\pm5.3}$	78.3$_{\pm5.3}$	79.3$_{\pm7.5}$	78.1$_{\pm7.9}$	46.9$_{\pm9.6}$	43.6$_{\pm10.4}$	71.2$_{\pm2.8}$	71.2$_{\pm2.8}$	54.2$_{\pm3.7}$	46.4$_{\pm8.3}$
	69.4$_{\pm6.2}$	66.2$_{\pm4.5}$	67.2$_{\pm4.2}$	61.1$_{\pm6.4}$	91.4$_{\pm2.1}$	88.3$_{\pm3.0}$	61.1$_{\pm6.9}$	54.4$_{\pm7.0}$	79.9$_{\pm1.2}$	78.5$_{\pm1.3}$	79.7$_{\pm7.4}$	79.6$_{\pm7.4}$	76.8$_{\pm9.7}$	76.2$_{\pm9.5}$	44.6$_{\pm10.4}$	35.0$_{\pm13.0}$	71.7$_{\pm2.3}$	71.6$_{\pm2.3}$	60.1$_{\pm2.1}$	57.8$_{\pm4.2}$
	7.7% ↓	8.3% ↓	1.3% ↓	5.0% ↓	0.4% ↓	0.5% ↓	3.6% ↑	9.6% ↑	0.8% ↓	0.6% ↓	1.6% ↑	1.6% ↑	3.1% ↓	2.4% ↓	4.8% ↓	19.7% ↓	0.6% ↑	0.6% ↑	10.7% ↑	24.4% ↑
Support	72.0$_{\pm8.6}$	69.3$_{\pm6.7}$	67.5$_{\pm4.1}$	63.7$_{\pm5.3}$	91.5$_{\pm1.3}$	88.6$_{\pm1.8}$	59.7$_{\pm7.3}$	49.2$_{\pm9.3}$	80.4$_{\pm1.1}$	78.9$_{\pm1.3}$	75.7$_{\pm8.6}$	75.4$_{\pm8.8}$	79.3$_{\pm7.9}$	78.5$_{\pm7.6}$	47.8$_{\pm6.0}$	44.7$_{\pm7.0}$	72.1$_{\pm2.2}$	72.0$_{\pm2.2}$	54.9$_{\pm3.3}$	47.6$_{\pm8.0}$
	72.3$_{\pm8.8}$	69.1$_{\pm8.4}$	66.8$_{\pm4.8}$	60.8$_{\pm7.3}$	91.3$_{\pm1.5}$	88.2$_{\pm2.0}$	61.8$_{\pm7.7}$	54.3$_{\pm9.2}$	80.1$_{\pm1.4}$	78.7$_{\pm1.5}$	79.0$_{\pm7.7}$	78.8$_{\pm7.7}$	77.9$_{\pm8.7}$	77.4$_{\pm8.6}$	46.4$_{\pm10.3}$	36.8$_{\pm12.6}$	71.9$_{\pm2.5}$	71.8$_{\pm2.6}$	60.2$_{\pm2.1}$	58.1$_{\pm4.1}$
	0.4% ↑	0.3% ↓	1.0% ↓	4.7% ↓	0.3% ↓	0.5% ↓	3.6% ↑	10.4% ↑	0.4% ↓	0.2% ↓	4.4% ↑	4.4% ↑	1.7% ↓	1.4% ↓	2.9% ↓	17.5% ↓	0.3% ↓	0.3% ↓	9.6% ↑	22.1% ↑
Oppose	73.7$_{\pm5.6}$	70.5$_{\pm4.8}$	67.2$_{\pm3.9}$	64.6$_{\pm5.0}$	90.8$_{\pm1.8}$	88.1$_{\pm2.2}$	60.7$_{\pm6.0}$	49.9$_{\pm9.4}$	80.0$_{\pm1.1}$	78.8$_{\pm1.1}$	77.9$_{\pm7.2}$	77.8$_{\pm7.3}$	79.0$_{\pm9.9}$	78.4$_{\pm9.9}$	48.9$_{\pm8.8}$	45.2$_{\pm10.1}$	70.8$_{\pm2.2}$	70.7$_{\pm2.2}$	54.4$_{\pm3.5}$	46.0$_{\pm8.1}$
	70.3$_{\pm9.5}$	66.1$_{\pm10.1}$	67.4$_{\pm4.9}$	61.7$_{\pm7.7}$	91.0$_{\pm1.5}$	88.2$_{\pm2.0}$	62.8$_{\pm7.2}$	54.7$_{\pm8.8}$	79.7$_{\pm1.2}$	78.4$_{\pm1.3}$	79.2$_{\pm7.2}$	79.0$_{\pm7.2}$	76.8$_{\pm8.2}$	76.1$_{\pm8.3}$	46.4$_{\pm10.6}$	37.4$_{\pm12.3}$	71.0$_{\pm2.3}$	70.9$_{\pm2.3}$	60.5$_{\pm2.1}$	58.4$_{\pm4.1}$
	4.6% ↓	6.2% ↓	0.3% ↑	4.5% ↓	0.3% ↑	0.1% ↑	3.5% ↑	9.6% ↑	0.4% ↓	0.5% ↓	1.6% ↑	1.6% ↑	2.8% ↓	2.9% ↓	5.1% ↓	17.4% ↓	0.3% ↑	0.3% ↑	11.3% ↑	26.7% ↑
Makeup	72.2$_{\pm9.3}$	69.6$_{\pm7.5}$	68.3$_{\pm3.0}$	65.0$_{\pm4.1}$	91.9$_{\pm1.6}$	89.0$_{\pm2.1}$	60.1$_{\pm5.7}$	48.6$_{\pm4.8}$	80.4$_{\pm1.3}$	78.9$_{\pm1.6}$	76.6$_{\pm7.4}$	76.4$_{\pm7.6}$	82.0$_{\pm6.5}$	81.0$_{\pm6.9}$	46.2$_{\pm7.5}$	42.5$_{\pm8.0}$	71.0$_{\pm1.7}$	70.9$_{\pm1.7}$	54.4$_{\pm4.2}$	45.8$_{\pm9.6}$
	73.7$_{\pm6.6}$	70.5$_{\pm5.1}$	67.4$_{\pm4.9}$	61.4$_{\pm7.4}$	91.3$_{\pm1.9}$	88.2$_{\pm2.5}$	64.3$_{\pm6.7}$	56.6$_{\pm8.0}$	80.1$_{\pm1.5}$	78.8$_{\pm1.7}$	79.7$_{\pm8.2}$	79.5$_{\pm8.2}$	79.3$_{\pm7.0}$	78.6$_{\pm7.1}$	46.4$_{\pm10.4}$	37.8$_{\pm11.4}$	71.4$_{\pm2.7}$	71.3$_{\pm2.8}$	60.4$_{\pm2.2}$	58.1$_{\pm4.3}$
	2.0% ↑	1.3% ↑	1.4% ↓	5.5% ↓	0.6% ↓	1.0% ↓	7.0% ↑	16.4% ↑	0.4% ↓	0.2% ↓	4.1% ↑	4.1% ↑	3.3% ↓	3.0% ↓	0.5% ↑	11.1% ↓	0.5% ↑	0.5% ↑	11.0% ↑	26.8% ↑
Amplify	71.8$_{\pm9.6}$	68.5$_{\pm8.2}$	67.1$_{\pm5.5}$	62.7$_{\pm6.7}$	91.5$_{\pm1.7}$	88.2$_{\pm2.3}$	60.2$_{\pm6.5}$	52.0$_{\pm8.2}$	80.6$_{\pm0.9}$	79.1$_{\pm1.1}$	75.9$_{\pm6.7}$	75.8$_{\pm6.7}$	79.3$_{\pm6.9}$	78.3$_{\pm6.9}$	51.3$_{\pm5.7}$	46.0$_{\pm7.6}$	70.5$_{\pm2.0}$	70.4$_{\pm2.0}$	54.4$_{\pm3.4}$	47.6$_{\pm8.1}$
	68.2$_{\pm9.3}$	68.2$_{\pm9.3}$	66.8$_{\pm5.0}$	60.3$_{\pm7.5}$	91.5$_{\pm2.0}$	88.4$_{\pm2.8}$	63.2$_{\pm7.4}$	57.8$_{\pm7.7}$	80.2$_{\pm1.2}$	78.8$_{\pm1.3}$	79.9$_{\pm6.8}$	79.7$_{\pm6.8}$	76.0$_{\pm9.3}$	75.2$_{\pm9.3}$	47.5$_{\pm10.9}$	37.4$_{\pm13.6}$	71.1$_{\pm2.3}$	70.9$_{\pm2.4}$	59.7$_{\pm1.9}$	57.5$_{\pm3.8}$
	5.0% ↓	0.5% ↓	0.4% ↓	3.8% ↓	0.0% ↑	0.2% ↑	4.9% ↑	11.0% ↑	0.5% ↓	0.4% ↓	5.3% ↑	5.1% ↑	4.2% ↓	4.0% ↓	7.4% ↓	18.7% ↓	0.8% ↑	0.7% ↑	9.8% ↑	20.8% ↑
Rephrase	75.6$_{\pm8.6}$	71.0$_{\pm9.2}$	68.8$_{\pm4.0}$	66.8$_{\pm4.3}$	91.3$_{\pm1.3}$	88.3$_{\pm1.8}$	63.2$_{\pm6.2}$	56.7$_{\pm6.3}$	80.6$_{\pm0.9}$	79.0$_{\pm1.1}$	79.2$_{\pm4.7}$	79.1$_{\pm4.8}$	82.8$_{\pm6.4}$	81.4$_{\pm7.8}$	49.1$_{\pm7.7}$	45.8$_{\pm7.7}$	72.3$_{\pm1.8}$	72.3$_{\pm1.8}$	65.3$_{\pm2.5}$	65.1$_{\pm2.5}$
	74.2$_{\pm6.7}$	70.2$_{\pm7.5}$	67.6$_{\pm4.8}$	61.8$_{\pm7.1}$	91.3$_{\pm1.6}$	88.2$_{\pm2.1}$	64.3$_{\pm6.6}$	58.1$_{\pm7.4}$	79.8$_{\pm1.4}$	78.5$_{\pm1.5}$	80.3$_{\pm6.2}$	80.1$_{\pm6.2}$	77.9$_{\pm8.4}$	77.2$_{\pm8.5}$	44.8$_{\pm11.0}$	36.3$_{\pm12.5}$	71.3$_{\pm1.7}$	71.2$_{\pm1.8}$	59.9$_{\pm2.6}$	57.4$_{\pm4.8}$
	1.9% ↓	1.1% ↓	1.9% ↓	7.6% ↓	0.0% ↑	0.1% ↓	1.8% ↑	2.5% ↑	1.0% ↓	0.7% ↓	1.4% ↑	1.3% ↑	6.0% ↓	5.1% ↓	8.8% ↓	20.7% ↓	1.5% ↓	1.5% ↓	8.3% ↓	11.9% ↓
Rewrite	72.8$_{\pm9.4}$	68.1$_{\pm10.0}$	68.6$_{\pm4.5}$	66.7$_{\pm5.2}$	92.3$_{\pm1.4}$	89.5$_{\pm1.9}$	62.4$_{\pm8.4}$	55.6$_{\pm9.4}$	80.5$_{\pm1.4}$	78.8$_{\pm1.7}$	78.1$_{\pm6.5}$	78.0$_{\pm6.5}$	85.6$_{\pm5.5}$	84.2$_{\pm7.2}$	47.0$_{\pm7.7}$	43.8$_{\pm8.3}$	72.4$_{\pm2.3}$	72.3$_{\pm2.4}$	65.6$_{\pm2.0}$	65.4$_{\pm2.1}$
	75.4$_{\pm7.2}$	71.7$_{\pm7.0}$	67.4$_{\pm5.1}$	61.4$_{\pm7.8}$	91.4$_{\pm1.4}$	88.2$_{\pm2.0}$	62.5$_{\pm8.3}$	55.5$_{\pm8.8}$	79.8$_{\pm1.5}$	78.4$_{\pm1.7}$	81.2$_{\pm6.8}$	81.0$_{\pm6.8}$	81.8$_{\pm9.2}$	81.1$_{\pm9.5}$	42.8$_{\pm12.0}$	32.4$_{\pm14.5}$	71.2$_{\pm2.4}$	71.1$_{\pm2.5}$	59.8$_{\pm2.4}$	57.2$_{\pm4.7}$
	3.7% ↑	5.4% ↑	1.8% ↓	7.9% ↓	0.9% ↓	1.4% ↓	0.2% ↑	0.2% ↓	0.8% ↓	0.5% ↓	4.0% ↑	3.9% ↑	4.5% ↓	3.8% ↓	9.1% ↓	26.1% ↓	1.6% ↓	1.7% ↓	8.8% ↓	12.6% ↓
Modify	73.0$_{\pm9.7}$	68.6$_{\pm8.6}$	69.8$_{\pm3.8}$	68.0$_{\pm4.3}$	92.1$_{\pm1.4}$	89.4$_{\pm1.9}$	62.5$_{\pm6.4}$	55.4$_{\pm7.3}$	80.6$_{\pm1.7}$	79.0$_{\pm2.1}$	77.9$_{\pm6.9}$	77.8$_{\pm6.9}$	83.1$_{\pm5.8}$	81.3$_{\pm8.8}$	47.3$_{\pm8.2}$	44.3$_{\pm7.3}$	71.9$_{\pm1.9}$	71.8$_{\pm1.9}$	65.0$_{\pm2.5}$	64.8$_{\pm2.6}$
	75.2$_{\pm6.9}$	71.6$_{\pm5.8}$	67.5$_{\pm4.7}$	61.7$_{\pm7.3}$	91.8$_{\pm1.5}$	88.8$_{\pm2.1}$	63.6$_{\pm6.2}$	56.5$_{\pm7.4}$	79.9$_{\pm1.6}$	78.6$_{\pm1.7}$	79.9$_{\pm6.8}$	79.8$_{\pm6.9}$	78.7$_{\pm8.7}$	78.0$_{\pm9.1}$	44.3$_{\pm11.8}$	34.3$_{\pm14.2}$	71.7$_{\pm1.8}$	71.7$_{\pm1.9}$	59.9$_{\pm2.3}$	57.3$_{\pm4.5}$
	3.0% ↑	4.3% ↑	3.3% ↓	9.3% ↓	0.4% ↓	0.7% ↓	1.8% ↑	2.1% ↑	0.8% ↓	0.6% ↓	2.6% ↑	2.6% ↑	5.3% ↓	4.0% ↓	6.2% ↓	22.6% ↓	0.3% ↓	0.2% ↓	7.8% ↓	11.5% ↓
Reverse	71.8$_{\pm8.8}$	67.4$_{\pm7.5}$	68.2$_{\pm3.8}$	66.9$_{\pm3.8}$	91.3$_{\pm1.4}$	88.5$_{\pm1.8}$	64.0$_{\pm8.2}$	56.7$_{\pm9.7}$	80.0$_{\pm1.3}$	78.2$_{\pm1.8}$	76.8$_{\pm4.6}$	76.7$_{\pm4.7}$	80.7$_{\pm5.0}$	79.3$_{\pm5.0}$	45.3$_{\pm8.9}$	41.8$_{\pm7.4}$	72.0$_{\pm2.7}$	71.9$_{\pm2.8}$	65.0$_{\pm2.1}$	64.8$_{\pm2.2}$
	74.4$_{\pm7.8}$	71.6$_{\pm7.4}$	67.1$_{\pm4.5}$	61.2$_{\pm7.2}$	90.9$_{\pm1.2}$	87.8$_{\pm1.5}$	63.6$_{\pm7.2}$	56.4$_{\pm7.8}$	79.9$_{\pm1.3}$	78.5$_{\pm1.5}$	80.3$_{\pm7.5}$	80.1$_{\pm7.5}$	80.1$_{\pm5.5}$	79.4$_{\pm5.5}$	44.3$_{\pm11.2}$	35.0$_{\pm14.6}$	71.7$_{\pm2.2}$	71.6$_{\pm2.3}$	60.1$_{\pm2.8}$	57.5$_{\pm5.0}$
	3.7% ↑	6.2% ↑	1.6% ↓	8.5% ↓	0.5% ↓	0.8% ↓	0.7% ↓	0.5% ↓	0.0% ↓	0.4% ↑	4.5% ↑	4.4% ↑	0.7% ↓	0.1% ↑	2.0% ↓	16.2% ↓	0.4% ↓	0.5% ↓	7.5% ↓	11.2% ↓
Remove	77.3$_{\pm4.5}$	73.1$_{\pm5.0}$	70.7$_{\pm3.5}$	67.8$_{\pm3.7}$	91.6$_{\pm1.3}$	88.5$_{\pm1.8}$	69.2$_{\pm4.7}$	66.2$_{\pm4.8}$	80.4$_{\pm1.1}$	78.9$_{\pm1.3}$	82.3$_{\pm4.9}$	82.2$_{\pm4.9}$	86.2$_{\pm8.2}$	85.3$_{\pm8.6}$	53.6$_{\pm7.4}$	49.6$_{\pm8.5}$	72.4$_{\pm2.2}$	72.4$_{\pm2.2}$	61.7$_{\pm2.1}$	60.2$_{\pm3.5}$
	72.8$_{\pm6.9}$	68.9$_{\pm7.5}$	68.1$_{\pm4.8}$	62.6$_{\pm7.2}$	91.9$_{\pm1.7}$	88.8$_{\pm2.3}$	68.1$_{\pm4.3}$	64.6$_{\pm4.6}$	79.7$_{\pm1.3}$	78.3$_{\pm1.4}$	82.1$_{\pm7.0}$	82.0$_{\pm7.1}$	82.6$_{\pm6.1}$	82.0$_{\pm6.1}$	47.7$_{\pm11.3}$	37.8$_{\pm12.6}$	71.6$_{\pm2.5}$	71.5$_{\pm2.7}$	59.9$_{\pm2.2}$	57.5$_{\pm4.0}$
	5.9% ↓	5.8% ↓	3.7% ↓	7.7% ↓	0.3% ↑	0.3% ↑	1.6% ↓	2.4% ↓	0.9% ↓	0.8% ↓	0.2% ↓	0.3% ↓	4.1% ↓	3.9% ↓	10.9% ↓	23.7% ↓	1.2% ↓	1.3% ↓	3.0% ↓	4.4% ↓
Repeat	72.3$_{\pm6.6}$	68.2$_{\pm7.1}$	67.3$_{\pm4.7}$	64.2$_{\pm5.2}$	90.7$_{\pm1.8}$	87.4$_{\pm2.5}$	66.0$_{\pm5.8}$	63.1$_{\pm5.3}$	79.6$_{\pm1.4}$	78.1$_{\pm1.6}$	79.2$_{\pm7.6}$	79.0$_{\pm7.7}$	77.1$_{\pm7.2}$	76.3$_{\pm7.2}$	46.4$_{\pm8.5}$	40.7$_{\pm9.6}$	70.2$_{\pm1.9}$	70.1$_{\pm2.0}$	55.1$_{\pm2.8}$	49.2$_{\pm6.5}$
	72.0$_{\pm5.3}$	68.2$_{\pm6.4}$	67.1$_{\pm4.7}$	61.5$_{\pm7.0}$	91.8$_{\pm1.9}$	88.8$_{\pm2.5}$	68.4$_{\pm5.0}$	65.3$_{\pm5.1}$	79.8$_{\pm1.3}$	78.4$_{\pm1.4}$	81.7$_{\pm6.9}$	81.6$_{\pm7.0}$	81.5$_{\pm9.3}$	80.9$_{\pm9.2}$	47.3$_{\pm7.8}$	40.1$_{\pm10.2}$	71.4$_{\pm2.4}$	71.3$_{\pm2.5}$	60.4$_{\pm2.5}$	58.2$_{\pm4.3}$
	0.3% ↓	0.1% ↑	0.3% ↓	4.2% ↓	1.2% ↑	1.5% ↑	3.6% ↑	3.5% ↑	0.3% ↑	0.4% ↑	3.3% ↑	3.3% ↑	5.7% ↑	6.1% ↑	1.9% ↑	1.5% ↓	1.7% ↑	1.7% ↑	9.7% ↑	18.3% ↑

Table 10: The **Mixture of Experts** strategy performance on DEFEND. We highlight the improved parts.

Pollution	Fake News						Hate Speech		Rumor								Sarcasm			
	Politifact		Gossipcop		ANTiVax		HASOC		Pheme		Twitter15		Twitter16		RumorEval		Twitter		Reddit	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
Vanilla	89.8$_{\pm6.2}$	88.5$_{\pm6.8}$	74.6$_{\pm2.4}$	72.9$_{\pm3.2}$	93.2$_{\pm0.8}$	90.7$_{\pm0.9}$	65.0$_{\pm5.1}$	61.6$_{\pm4.8}$	85.5$_{\pm1.6}$	84.5$_{\pm1.7}$	91.3$_{\pm2.5}$	91.3$_{\pm2.4}$	94.2$_{\pm3.3}$	93.9$_{\pm3.6}$	60.8$_{\pm4.7}$	59.6$_{\pm4.8}$	73.7$_{\pm2.0}$	73.6$_{\pm2.0}$	65.1$_{\pm2.0}$	64.8$_{\pm2.2}$
	87.4$_{\pm7.4}$	85.9$_{\pm7.5}$	74.5$_{\pm2.2}$	72.5$_{\pm3.2}$	93.2$_{\pm0.8}$	90.8$_{\pm0.9}$	64.6$_{\pm5.3}$	61.1$_{\pm5.2}$	85.4$_{\pm1.6}$	84.5$_{\pm1.7}$	91.3$_{\pm3.0}$	91.3$_{\pm3.0}$	94.2$_{\pm2.7}$	93.9$_{\pm3.0}$	61.4$_{\pm4.9}$	60.4$_{\pm5.1}$	73.5$_{\pm1.8}$	73.4$_{\pm1.8}$	65.7$_{\pm2.3}$	65.2$_{\pm2.7}$
	2.7% ↓	2.9% ↓	0.1% ↓	0.5% ↓	0.0% ↑	0.1% ↑	0.6% ↓	0.8% ↓	0.0% ↓	0.0% ↑	0.0% ↑	0.0% ↓	0.0% ↑	0.0% ↑	1.1% ↑	1.3% ↑	0.2% ↓	0.2% ↓	0.8% ↑	0.7% ↑
Publisher	89.1$_{\pm8.3}$	87.9$_{\pm8.5}$	74.8$_{\pm2.7}$	73.2$_{\pm2.9}$	93.2$_{\pm0.6}$	90.7$_{\pm0.8}$	65.5$_{\pm4.4}$	61.8$_{\pm4.6}$	85.4$_{\pm1.5}$	84.4$_{\pm1.6}$	91.0$_{\pm2.8}$	90.9$_{\pm2.8}$	94.2$_{\pm3.8}$	93.9$_{\pm4.0}$	61.4$_{\pm4.7}$	60.4$_{\pm4.7}$	73.6$_{\pm1.5}$	73.5$_{\pm1.5}$	65.4$_{\pm1.7}$	65.0$_{\pm1.9}$
	86.5$_{\pm6.8}$	84.6$_{\pm7.4}$	75.0$_{\pm2.0}$	73.1$_{\pm2.8}$	93.3$_{\pm0.7}$	90.9$_{\pm0.7}$	65.5$_{\pm4.3}$	61.6$_{\pm5.2}$	85.3$_{\pm1.5}$	84.4$_{\pm1.6}$	91.0$_{\pm2.8}$	90.9$_{\pm2.8}$	94.2$_{\pm3.8}$	93.9$_{\pm4.0}$	60.3$_{\pm5.0}$	59.3$_{\pm5.0}$	73.6$_{\pm1.6}$	73.6$_{\pm1.6}$	65.4$_{\pm2.5}$	65.0$_{\pm2.8}$
	3.0% ↓	3.8% ↓	0.3% ↑	0.1% ↓	0.1% ↑	0.1% ↑	0.0% ↑	0.2% ↓	0.1% ↓	0.0% ↑	0.0% ↑	0.0% ↑	0.0% ↑	0.0% ↑	1.8% ↓	1.8% ↓	0.1% ↑	0.1% ↑	0.1% ↑	0.0% ↑
Echo	90.3$_{\pm6.0}$	88.9$_{\pm6.2}$	74.6$_{\pm2.6}$	72.9$_{\pm2.9}$	93.3$_{\pm0.6}$	90.9$_{\pm0.6}$	66.0$_{\pm6.2}$	62.5$_{\pm6.6}$	85.6$_{\pm1.4}$	84.7$_{\pm1.6}$	91.0$_{\pm2.8}$	90.9$_{\pm2.8}$	94.2$_{\pm3.0}$	93.9$_{\pm3.3}$	60.5$_{\pm5.2}$	59.5$_{\pm5.7}$	73.7$_{\pm1.8}$	73.7$_{\pm1.8}$	65.1$_{\pm2.1}$	64.7$_{\pm2.3}$
	87.4$_{\pm6.5}$	85.8$_{\pm6.5}$	74.8$_{\pm2.6}$	72.7$_{\pm3.5}$	93.2$_{\pm0.9}$	90.8$_{\pm1.1}$	65.6$_{\pm5.7}$	61.8$_{\pm6.8}$	85.5$_{\pm1.4}$	84.7$_{\pm1.5}$	91.0$_{\pm2.8}$	90.9$_{\pm2.8}$	94.2$_{\pm3.0}$	93.9$_{\pm3.3}$	59.6$_{\pm5.9}$	58.8$_{\pm6.3}$	73.5$_{\pm1.9}$	73.5$_{\pm1.9}$	65.4$_{\pm2.0}$	65.0$_{\pm2.4}$
	3.2% ↓	3.6% ↓	0.3% ↑	0.3% ↓	0.1% ↓	0.1% ↓	0.6% ↓	1.2% ↓	0.1% ↓	0.1% ↓	0.0% ↑	0.0% ↑	0.0% ↑	0.0% ↑	1.5% ↓	1.1% ↓	0.2% ↓	0.2% ↓	0.5% ↑	0.4% ↑
Support	89.6$_{\pm6.0}$	88.2$_{\pm6.4}$	74.0$_{\pm2.8}$	72.4$_{\pm3.4}$	93.3$_{\pm0.6}$	90.8$_{\pm0.7}$	65.2$_{\pm6.0}$	61.6$_{\pm6.1}$	85.6$_{\pm1.4}$	84.7$_{\pm1.5}$	91.3$_{\pm3.0}$	91.3$_{\pm3.0}$	93.7$_{\pm3.9}$	93.3$_{\pm4.3}$	60.8$_{\pm6.4}$	59.8$_{\pm6.6}$	73.5$_{\pm1.8}$	73.4$_{\pm1.8}$	65.5$_{\pm1.8}$	65.1$_{\pm2.0}$
	85.7$_{\pm8.1}$	83.7$_{\pm8.7}$	74.6$_{\pm2.6}$	72.7$_{\pm3.6}$	93.3$_{\pm0.7}$	90.9$_{\pm0.8}$	66.4$_{\pm4.7}$	63.0$_{\pm5.3}$	85.5$_{\pm1.4}$	84.6$_{\pm1.5}$	91.3$_{\pm3.0}$	91.3$_{\pm3.0}$	93.4$_{\pm4.1}$	93.1$_{\pm4.5}$	61.6$_{\pm5.9}$	60.7$_{\pm5.9}$	73.7$_{\pm1.8}$	73.7$_{\pm1.8}$	65.5$_{\pm1.7}$	65.0$_{\pm2.1}$
	4.3% ↓	5.1% ↓	0.8% ↑	0.4% ↑	0.0% ↑	0.1% ↑	1.9% ↑	2.2% ↑	0.2% ↓	0.1% ↓	0.0% ↑	0.0% ↑	0.3% ↓	0.3% ↓	1.5% ↑	1.4% ↑	0.4% ↑	0.3% ↑	0.1% ↓	0.2% ↓
Oppose	90.3$_{\pm4.6}$	88.9$_{\pm5.0}$	74.6$_{\pm2.1}$	73.0$_{\pm2.5}$	93.3$_{\pm0.6}$	90.8$_{\pm0.8}$	65.9$_{\pm5.3}$	62.1$_{\pm5.9}$	85.6$_{\pm1.6}$	84.6$_{\pm1.7}$	91.2$_{\pm3.0}$	91.1$_{\pm3.0}$	94.2$_{\pm3.8}$	93.9$_{\pm4.0}$	61.9$_{\pm5.9}$	60.9$_{\pm5.7}$	73.8$_{\pm1.7}$	73.7$_{\pm1.7}$	65.8$_{\pm2.3}$	65.4$_{\pm2.5}$
	86.7$_{\pm9.9}$	84.3$_{\pm11.3}$	74.7$_{\pm2.2}$	72.8$_{\pm3.2}$	93.3$_{\pm0.8}$	90.9$_{\pm0.9}$	67.3$_{\pm5.2}$	63.9$_{\pm5.3}$	85.6$_{\pm1.4}$	84.7$_{\pm1.5}$	91.2$_{\pm3.0}$	91.1$_{\pm3.0}$	94.2$_{\pm3.8}$	93.9$_{\pm4.0}$	62.1$_{\pm6.1}$	61.2$_{\pm6.0}$	73.8$_{\pm1.8}$	73.8$_{\pm1.8}$	65.5$_{\pm1.9}$	65.0$_{\pm2.3}$
	4.0% ↓	5.2% ↓	0.2% ↑	0.2% ↓	0.0% ↑	0.1% ↑	2.1% ↑	2.9% ↑	0.0% ↑	0.1% ↑	0.0% ↑	0.0% ↑	0.0% ↑	0.0% ↑	0.4% ↑	0.5% ↑	0.0% ↑	0.0% ↑	0.4% ↓	0.5% ↓
Makeup	90.8$_{\pm6.2}$	89.6$_{\pm6.2}$	74.6$_{\pm2.3}$	72.9$_{\pm3.1}$	93.2$_{\pm0.7}$	90.7$_{\pm1.0}$	66.6$_{\pm5.9}$	63.1$_{\pm7.2}$	85.3$_{\pm1.4}$	84.4$_{\pm1.5}$	91.2$_{\pm3.4}$	91.1$_{\pm3.4}$	94.2$_{\pm3.0}$	93.9$_{\pm3.3}$	60.3$_{\pm5.8}$	59.2$_{\pm5.6}$	73.6$_{\pm1.6}$	73.5$_{\pm1.6}$	65.9$_{\pm2.2}$	65.5$_{\pm2.4}$
	87.7$_{\pm6.7}$	85.3$_{\pm8.6}$	74.5$_{\pm2.2}$	72.5$_{\pm3.3}$	93.3$_{\pm0.8}$	90.9$_{\pm0.8}$	66.3$_{\pm5.1}$	62.7$_{\pm5.5}$	85.5$_{\pm1.3}$	84.7$_{\pm1.3}$	91.2$_{\pm2.8}$	91.1$_{\pm2.8}$	93.7$_{\pm3.2}$	93.4$_{\pm3.5}$	61.0$_{\pm6.3}$	59.9$_{\pm6.3}$	73.8$_{\pm1.5}$	73.7$_{\pm1.5}$	65.4$_{\pm1.9}$	65.0$_{\pm2.2}$
	3.4% ↓	4.8% ↓	0.2% ↓	0.6% ↓	0.2% ↑	0.2% ↑	0.4% ↓	0.5% ↓	0.2% ↑	0.3% ↑	0.0% ↑	0.0% ↑	0.6% ↓	0.6% ↓	1.1% ↑	1.3% ↑	0.2% ↑	0.2% ↑	0.7% ↓	0.8% ↓
Amplify	89.8$_{\pm6.9}$	88.4$_{\pm7.2}$	74.9$_{\pm2.1}$	73.3$_{\pm2.4}$	93.4$_{\pm0.7}$	91.1$_{\pm0.8}$	65.2$_{\pm5.7}$	61.4$_{\pm6.8}$	85.4$_{\pm1.5}$	84.5$_{\pm1.6}$	91.2$_{\pm2.8}$	91.1$_{\pm2.8}$	93.9$_{\pm4.1}$	93.6$_{\pm4.4}$	60.5$_{\pm6.1}$	59.0$_{\pm6.0}$	73.8$_{\pm1.7}$	73.7$_{\pm1.7}$	65.3$_{\pm2.1}$	64.9$_{\pm2.3}$
	86.7$_{\pm7.1}$	84.8$_{\pm7.3}$	74.9$_{\pm2.3}$	73.0$_{\pm2.8}$	93.2$_{\pm0.8}$	90.8$_{\pm0.9}$	64.2$_{\pm5.3}$	60.5$_{\pm5.2}$	85.4$_{\pm1.5}$	84.5$_{\pm1.6}$	91.0$_{\pm2.8}$	90.9$_{\pm2.8}$	93.6$_{\pm4.5}$	93.3$_{\pm4.9}$	61.6$_{\pm6.0}$	60.5$_{\pm5.9}$	73.6$_{\pm1.8}$	73.5$_{\pm1.8}$	65.5$_{\pm2.1}$	65.0$_{\pm2.6}$
	3.5% ↓	4.1% ↓	0.1% ↑	0.4% ↓	0.2% ↓	0.3% ↓	1.5% ↓	1.4% ↓	0.0% ↓	0.0% ↑	0.2% ↓	0.2% ↓	0.3% ↓	0.3% ↓	1.8% ↑	2.6% ↑	0.2% ↓	0.2% ↓	0.2% ↑	0.0% ↑
Rephrase	91.5$_{\pm4.5}$	90.5$_{\pm4.5}$	74.7$_{\pm2.0}$	73.1$_{\pm2.6}$	93.4$_{\pm0.7}$	91.0$_{\pm0.8}$	66.6$_{\pm4.9}$	63.7$_{\pm4.9}$	85.4$_{\pm1.7}$	84.4$_{\pm1.8}$	91.0$_{\pm3.5}$	90.9$_{\pm3.5}$	93.6$_{\pm2.9}$	93.3$_{\pm3.2}$	62.8$_{\pm4.3}$	61.5$_{\pm4.6}$	73.6$_{\pm1.8}$	73.5$_{\pm1.8}$	65.6$_{\pm2.4}$	65.2$_{\pm2.6}$
	87.5$_{\pm4.5}$	85.0$_{\pm6.7}$	74.4$_{\pm2.6}$	72.2$_{\pm4.3}$	93.3$_{\pm0.9}$	91.0$_{\pm1.0}$	65.3$_{\pm4.9}$	62.1$_{\pm4.8}$	85.2$_{\pm1.6}$	84.3$_{\pm1.7}$	90.8$_{\pm3.4}$	90.7$_{\pm3.4}$	93.4$_{\pm2.7}$	93.1$_{\pm2.9}$	61.9$_{\pm4.8}$	60.7$_{\pm5.0}$	73.5$_{\pm1.7}$	73.5$_{\pm1.7}$	65.6$_{\pm2.0}$	65.0$_{\pm2.3}$
	4.5% ↓	6.1% ↓	0.4% ↓	1.3% ↓	0.1% ↓	0.1% ↓	1.9% ↓	2.5% ↓	0.3% ↓	0.2% ↓	0.2% ↓	0.2% ↓	0.3% ↓	0.3% ↓	1.5% ↓	1.3% ↓	0.1% ↓	0.1% ↓	0.0% ↑	0.3% ↓
Rewrite	92.0$_{\pm5.6}$	90.8$_{\pm6.0}$	74.7$_{\pm2.6}$	73.2$_{\pm3.0}$	93.0$_{\pm0.7}$	90.5$_{\pm0.9}$	65.9$_{\pm5.0}$	62.7$_{\pm5.0}$	85.4$_{\pm1.5}$	84.5$_{\pm1.6}$	91.2$_{\pm3.6}$	91.1$_{\pm3.6}$	94.2$_{\pm3.8}$	93.9$_{\pm4.1}$	61.9$_{\pm6.1}$	60.7$_{\pm6.1}$	73.7$_{\pm2.1}$	73.6$_{\pm2.1}$	65.6$_{\pm1.9}$	65.2$_{\pm2.2}$
	88.7$_{\pm7.2}$	86.1$_{\pm9.5}$	74.7$_{\pm2.5}$	72.6$_{\pm3.9}$	93.2$_{\pm0.8}$	90.7$_{\pm0.9}$	65.6$_{\pm4.3}$	62.3$_{\pm4.1}$	85.3$_{\pm1.5}$	84.4$_{\pm1.6}$	91.0$_{\pm3.4}$	90.9$_{\pm3.4}$	94.2$_{\pm3.8}$	93.9$_{\pm4.1}$	61.2$_{\pm6.3}$	60.1$_{\pm6.5}$	73.6$_{\pm1.6}$	73.6$_{\pm1.6}$	65.6$_{\pm2.1}$	65.0$_{\pm2.6}$
	3.7% ↓	5.1% ↓	0.1% ↓	0.8% ↓	0.2% ↑	0.3% ↑	0.4% ↓	0.7% ↓	0.1% ↓	0.1% ↓	0.2% ↓	0.2% ↓	0.0% ↑	0.0% ↑	1.1% ↓	0.9% ↓	0.1% ↓	0.1% ↓	0.0% ↑	0.4% ↓
Modify	91.8$_{\pm6.2}$	90.6$_{\pm6.6}$	75.0$_{\pm2.0}$	73.5$_{\pm2.2}$	93.3$_{\pm0.7}$	90.9$_{\pm0.7}$	66.0$_{\pm6.8}$	63.0$_{\pm7.0}$	85.6$_{\pm1.2}$	84.6$_{\pm1.2}$	91.3$_{\pm3.2}$	91.3$_{\pm3.2}$	93.7$_{\pm3.2}$	93.3$_{\pm3.5}$	61.7$_{\pm4.2}$	60.6$_{\pm4.6}$	73.7$_{\pm1.8}$	73.7$_{\pm1.8}$	65.5$_{\pm2.3}$	65.1$_{\pm2.5}$
	89.6$_{\pm5.8}$	87.4$_{\pm7.0}$	74.3$_{\pm2.8}$	72.2$_{\pm4.0}$	93.2$_{\pm0.8}$	90.8$_{\pm0.9}$	65.7$_{\pm7.2}$	62.5$_{\pm7.5}$	85.3$_{\pm1.6}$	84.4$_{\pm1.7}$	91.0$_{\pm2.9}$	90.9$_{\pm2.9}$	93.4$_{\pm3.7}$	93.1$_{\pm4.1}$	60.3$_{\pm5.5}$	59.2$_{\pm5.5}$	73.4$_{\pm1.8}$	73.4$_{\pm1.8}$	65.7$_{\pm2.1}$	65.1$_{\pm2.5}$
	2.3% ↓	3.5% ↓	1.0% ↓	1.7% ↓	0.1% ↓	0.2% ↓	0.4% ↓	0.7% ↓	0.3% ↓	0.2% ↓	0.4% ↓	0.4% ↓	0.3% ↓	0.3% ↓	2.2% ↓	2.2% ↓	0.4% ↓	0.4% ↓	0.4% ↑	0.0% ↑
Reverse	92.0$_{\pm6.0}$	90.9$_{\pm6.6}$	75.0$_{\pm2.4}$	73.5$_{\pm2.7}$	93.4$_{\pm0.7}$	91.0$_{\pm0.7}$	67.0$_{\pm6.6}$	63.8$_{\pm6.8}$	85.7$_{\pm1.5}$	84.8$_{\pm1.6}$	91.2$_{\pm3.1}$	91.1$_{\pm3.1}$	93.9$_{\pm3.4}$	93.6$_{\pm3.7}$	61.7$_{\pm4.5}$	60.4$_{\pm4.6}$	73.7$_{\pm1.6}$	73.7$_{\pm1.6}$	65.5$_{\pm1.9}$	65.1$_{\pm2.1}$
	89.6$_{\pm6.9}$	87.9$_{\pm7.3}$	75.2$_{\pm2.5}$	73.4$_{\pm3.2}$	93.4$_{\pm0.7}$	91.0$_{\pm0.8}$	66.4$_{\pm6.5}$	63.1$_{\pm6.3}$	85.3$_{\pm1.6}$	84.4$_{\pm1.7}$	91.0$_{\pm3.0}$	90.9$_{\pm3.0}$	94.2$_{\pm3.0}$	93.9$_{\pm3.3}$	61.4$_{\pm5.2}$	60.2$_{\pm5.9}$	73.7$_{\pm1.6}$	73.7$_{\pm1.6}$	65.6$_{\pm2.3}$	65.0$_{\pm2.7}$
	2.6% ↓	3.3% ↓	0.3% ↑	0.1% ↓	0.0% ↓	0.0% ↓	0.8% ↓	1.0% ↓	0.5% ↓	0.4% ↓	0.2% ↓	0.2% ↓	0.3% ↑	0.3% ↑	0.4% ↓	0.3% ↓	0.0% ↑	0.0% ↑	0.1% ↑	0.2% ↓
Remove	92.3$_{\pm3.0}$	91.1$_{\pm3.1}$	75.2$_{\pm2.4}$	73.5$_{\pm2.7}$	93.5$_{\pm0.6}$	91.1$_{\pm0.7}$	68.0$_{\pm5.1}$	65.0$_{\pm5.8}$	85.7$_{\pm1.4}$	84.7$_{\pm1.5}$	92.3$_{\pm2.1}$	92.2$_{\pm2.1}$	94.2$_{\pm3.8}$	93.9$_{\pm4.1}$	63.0$_{\pm4.3}$	62.0$_{\pm5.0}$	73.8$_{\pm1.7}$	73.8$_{\pm1.7}$	65.9$_{\pm2.0}$	65.5$_{\pm2.3}$
	87.0$_{\pm5.8}$	84.9$_{\pm7.1}$	74.2$_{\pm2.5}$	72.4$_{\pm3.1}$	93.4$_{\pm0.7}$	91.0$_{\pm0.9}$	67.1$_{\pm5.5}$	64.0$_{\pm5.8}$	85.4$_{\pm1.6}$	84.5$_{\pm1.7}$	92.1$_{\pm2.3}$	92.0$_{\pm2.3}$	94.2$_{\pm3.3}$	93.9$_{\pm3.5}$	60.5$_{\pm5.8}$	59.0$_{\pm6.3}$	73.8$_{\pm1.4}$	73.7$_{\pm1.4}$	64.9$_{\pm2.2}$	64.3$_{\pm2.7}$
	5.8% ↓	6.8% ↓	1.3% ↓	1.6% ↓	0.1% ↓	0.1% ↓	1.2% ↓	1.4% ↓	0.3% ↓	0.2% ↓	0.2% ↓	0.2% ↓	0.0% ↓	0.0% ↑	4.0% ↓	4.8% ↓	0.1% ↓	0.1% ↓	1.5% ↓	1.7% ↓
Repeat	86.0$_{\pm5.6}$	83.9$_{\pm7.3}$	74.2$_{\pm2.2}$	72.4$_{\pm3.0}$	93.3$_{\pm0.8}$	90.9$_{\pm1.0}$	68.3$_{\pm5.4}$	65.5$_{\pm6.2}$	85.3$_{\pm1.7}$	84.4$_{\pm1.8}$	91.9$_{\pm2.6}$	91.8$_{\pm2.6}$	93.4$_{\pm3.5}$	93.1$_{\pm3.8}$	60.5$_{\pm5.9}$	59.1$_{\pm7.1}$	73.8$_{\pm1.5}$	73.7$_{\pm1.5}$	65.3$_{\pm2.3}$	64.9$_{\pm2.7}$
	85.3$_{\pm7.4}$	83.1$_{\pm8.4}$	74.4$_{\pm2.0}$	72.5$_{\pm2.9}$	93.3$_{\pm0.9}$	90.9$_{\pm1.2}$	67.1$_{\pm5.3}$	64.1$_{\pm5.6}$	85.3$_{\pm1.7}$	84.3$_{\pm1.8}$	92.3$_{\pm2.3}$	92.2$_{\pm2.2}$	93.4$_{\pm3.5}$	93.1$_{\pm3.8}$	60.1$_{\pm5.6}$	58.6$_{\pm6.9}$	73.8$_{\pm1.5}$	73.7$_{\pm1.5}$	65.5$_{\pm2.2}$	65.1$_{\pm2.6}$
	0.9% ↓	0.9% ↓	0.2% ↑	0.1% ↑	0.1% ↑	0.1% ↑	1.6% ↓	2.2% ↓	0.0% ↓	0.0% ↓	0.4% ↑	0.4% ↑	0.0% ↑	0.0% ↑	0.7% ↓	0.7% ↓	0.0% ↑	0.0% ↑	0.2% ↑	0.3% ↑

Table 11: The **Mixture of Experts** strategy performance on GET. We highlight the improved parts.

Pollution	Fake News						Hate Speech		Rumor								Sarcasm			
	Politifact		Gossipcop		ANTiVax		HASOC		Pheme		Twitter15		Twitter16		RumorEval		Twitter		Reddit	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
Vanilla	$79.3_{\pm 11.0}$ $78.5_{\pm 9.9}$ 0.9% ↓	$78.2_{\pm 10.4}$ $77.3_{\pm 9.4}$ 1.1% ↓	$71.2_{\pm 3.6}$ $71.0_{\pm 2.7}$ 0.2% ↓	$70.4_{\pm 3.6}$ $70.3_{\pm 2.7}$ 0.2% ↓	$93.0_{\pm 1.1}$ $93.0_{\pm 1.2}$ 0.0% ↑	$90.1_{\pm 2.1}$ $90.1_{\pm 2.1}$ 0.0% ↑	$64.8_{\pm 7.6}$ $64.6_{\pm 8.4}$ 0.2% ↓	$56.8_{\pm 7.9}$ $56.5_{\pm 8.9}$ 0.5% ↓	$85.2_{\pm 1.8}$ $85.1_{\pm 1.7}$ 0.1% ↓	$84.0_{\pm 1.8}$ $83.9_{\pm 1.7}$ 0.1% ↓	$86.2_{\pm 4.1}$ $86.6_{\pm 4.8}$ 0.4% ↑	$86.1_{\pm 4.1}$ $86.5_{\pm 4.9}$ 0.4% ↑	$76.8_{\pm 6.7}$ $76.5_{\pm 7.9}$ 0.4% ↓	$76.1_{\pm 6.9}$ $75.8_{\pm 8.1}$ 0.4% ↓	$58.8_{\pm 7.8}$ $59.2_{\pm 7.4}$ 0.8% ↑	$54.9_{\pm 8.2}$ $55.5_{\pm 7.6}$ 1.1% ↑	$74.3_{\pm 2.0}$ $74.0_{\pm 1.9}$ 0.4% ↓	$74.2_{\pm 2.0}$ $74.0_{\pm 1.9}$ 0.4% ↓	$65.6_{\pm 3.3}$ $65.7_{\pm 3.4}$ 0.2% ↑	$65.2_{\pm 3.5}$ $65.2_{\pm 3.7}$ 0.1% ↑
Publisher	$80.0_{\pm 12.1}$ $79.5_{\pm 10.3}$ 0.6% ↓	$78.9_{\pm 11.6}$ $78.3_{\pm 9.7}$ 0.7% ↓	$71.5_{\pm 2.1}$ $71.3_{\pm 2.4}$ 0.3% ↓	$69.8_{\pm 2.5}$ $69.4_{\pm 2.9}$ 0.5% ↓	$93.2_{\pm 1.1}$ $93.3_{\pm 1.2}$ 0.2% ↑	$90.4_{\pm 1.9}$ $90.6_{\pm 2.1}$ 0.2% ↑	$63.2_{\pm 7.0}$ $63.8_{\pm 6.8}$ 0.9% ↑	$52.2_{\pm 8.3}$ $53.5_{\pm 7.7}$ 2.5% ↑	$84.8_{\pm 1.3}$ $84.9_{\pm 1.4}$ 0.2% ↑	$83.5_{\pm 1.1}$ $83.7_{\pm 1.3}$ 0.2% ↑	$85.3_{\pm 4.9}$ $84.6_{\pm 5.0}$ 0.9% ↓	$85.1_{\pm 5.0}$ $84.4_{\pm 5.1}$ 0.8% ↓	$75.4_{\pm 6.5}$ $75.4_{\pm 6.6}$ 0.0% ↑	$74.5_{\pm 6.5}$ $74.5_{\pm 6.6}$ 0.0% ↑	$58.8_{\pm 7.3}$ $58.8_{\pm 7.4}$ 0.0% ↓	$54.0_{\pm 9.3}$ $53.8_{\pm 8.5}$ 0.5% ↓	$74.2_{\pm 2.3}$ $74.0_{\pm 2.0}$ 0.3% ↓	$74.2_{\pm 2.3}$ $73.9_{\pm 2.0}$ 0.4% ↓	$65.5_{\pm 3.2}$ $65.3_{\pm 3.3}$ 0.4% ↓	$65.2_{\pm 3.4}$ $64.9_{\pm 3.6}$ 0.5% ↓
Echo	$81.9_{\pm 9.5}$ $82.2_{\pm 9.7}$ 0.3% ↑	$81.0_{\pm 9.1}$ $81.2_{\pm 9.2}$ 0.3% ↑	$71.7_{\pm 2.2}$ $71.3_{\pm 2.7}$ 0.6% ↓	$69.7_{\pm 2.9}$ $69.0_{\pm 3.3}$ 0.9% ↓	$93.1_{\pm 1.4}$ $93.2_{\pm 1.4}$ 0.1% ↑	$90.2_{\pm 2.6}$ $90.3_{\pm 2.4}$ 0.1% ↑	$67.3_{\pm 7.1}$ $66.9_{\pm 7.0}$ 0.6% ↓	$61.8_{\pm 7.4}$ $61.4_{\pm 7.3}$ 0.8% ↓	$84.7_{\pm 1.5}$ $84.8_{\pm 1.4}$ 0.0% ↑	$83.3_{\pm 1.5}$ $83.4_{\pm 1.5}$ 0.0% ↑	$85.7_{\pm 5.2}$ $85.1_{\pm 4.8}$ 0.6% ↓	$85.5_{\pm 5.2}$ $85.0_{\pm 4.9}$ 0.7% ↓	$77.9_{\pm 6.8}$ $77.3_{\pm 7.0}$ 0.7% ↓	$77.2_{\pm 6.4}$ $76.7_{\pm 6.8}$ 0.7% ↓	$60.5_{\pm 5.9}$ $60.8_{\pm 8.3}$ 0.4% ↑	$56.3_{\pm 5.6}$ $56.9_{\pm 9.5}$ 1.1% ↑	$74.2_{\pm 2.7}$ $74.3_{\pm 2.4}$ 0.1% ↑	$74.0_{\pm 2.8}$ $74.1_{\pm 2.5}$ 0.2% ↑	$66.7_{\pm 2.6}$ $66.5_{\pm 2.4}$ 0.2% ↓	$66.5_{\pm 2.7}$ $66.4_{\pm 2.5}$ 0.2% ↓
Support	$79.3_{\pm 11.1}$ $78.8_{\pm 11.3}$ 0.6% ↓	$78.2_{\pm 10.5}$ $77.6_{\pm 10.8}$ 0.7% ↓	$70.9_{\pm 3.2}$ $70.9_{\pm 4.4}$ 0.0% ↑	$70.1_{\pm 3.3}$ $70.1_{\pm 4.4}$ 0.0% ↓	$93.3_{\pm 1.3}$ $93.4_{\pm 1.2}$ 0.2% ↑	$90.5_{\pm 2.5}$ $90.7_{\pm 2.4}$ 0.3% ↑	$63.8_{\pm 8.1}$ $64.2_{\pm 7.1}$ 0.7% ↑	$55.2_{\pm 8.4}$ $55.7_{\pm 7.5}$ 0.8% ↑	$85.0_{\pm 1.8}$ $84.9_{\pm 1.9}$ 0.1% ↓	$83.7_{\pm 1.8}$ $83.6_{\pm 1.8}$ 0.1% ↓	$86.0_{\pm 4.4}$ $85.8_{\pm 4.3}$ 0.2% ↓	$85.9_{\pm 4.5}$ $85.7_{\pm 4.3}$ 0.2% ↓	$76.8_{\pm 7.3}$ $75.7_{\pm 8.0}$ 1.4% ↓	$76.2_{\pm 7.4}$ $75.1_{\pm 8.0}$ 1.5% ↓	$60.3_{\pm 6.7}$ $60.6_{\pm 7.2}$ 0.4% ↑	$56.2_{\pm 7.6}$ $56.5_{\pm 7.8}$ 0.5% ↑	$74.9_{\pm 2.7}$ $75.0_{\pm 2.6}$ 0.1% ↑	$74.8_{\pm 2.7}$ $74.9_{\pm 2.5}$ 0.1% ↑	$65.9_{\pm 2.8}$ $65.8_{\pm 2.7}$ 0.1% ↓	$65.5_{\pm 2.9}$ $65.4_{\pm 2.9}$ 0.2% ↓
Oppose	$82.4_{\pm 11.2}$ $81.0_{\pm 12.0}$ 1.7% ↓	$81.5_{\pm 10.5}$ $80.0_{\pm 11.6}$ 1.8% ↓	$66.5_{\pm 3.9}$ $66.8_{\pm 3.5}$ 0.4% ↑	$66.2_{\pm 4.1}$ $66.5_{\pm 3.5}$ 0.6% ↑	$93.3_{\pm 1.5}$ $93.2_{\pm 1.4}$ 0.1% ↓	$91.0_{\pm 1.9}$ $90.9_{\pm 1.8}$ 0.2% ↓	$65.0_{\pm 8.3}$ $66.0_{\pm 8.4}$ 1.5% ↑	$57.1_{\pm 8.9}$ $58.8_{\pm 8.7}$ 2.9% ↑	$82.9_{\pm 2.3}$ $82.7_{\pm 2.5}$ 0.2% ↓	$82.2_{\pm 2.1}$ $82.0_{\pm 2.2}$ 0.2% ↓	$86.6_{\pm 4.8}$ $85.6_{\pm 5.9}$ 1.1% ↓	$86.4_{\pm 5.0}$ $85.4_{\pm 6.1}$ 1.1% ↓	$79.8_{\pm 7.5}$ $79.8_{\pm 7.2}$ 0.0% ↓	$79.4_{\pm 7.3}$ $79.3_{\pm 7.0}$ 0.0% ↓	$61.5_{\pm 5.7}$ $60.8_{\pm 6.6}$ 1.1% ↓	$59.7_{\pm 5.3}$ $58.8_{\pm 6.1}$ 1.5% ↓	$73.3_{\pm 2.7}$ $73.6_{\pm 2.4}$ 0.4% ↑	$73.1_{\pm 2.7}$ $73.5_{\pm 2.3}$ 0.5% ↑	$67.0_{\pm 2.8}$ $66.9_{\pm 2.8}$ 0.2% ↓	$66.8_{\pm 2.8}$ $66.7_{\pm 2.9}$ 0.2% ↓
Makeup	$80.5_{\pm 12.0}$ $79.8_{\pm 10.9}$ 0.9% ↓	$79.5_{\pm 11.5}$ $78.7_{\pm 10.4}$ 1.0% ↓	$71.2_{\pm 2.9}$ $70.8_{\pm 2.4}$ 0.5% ↓	$70.2_{\pm 3.1}$ $69.8_{\pm 2.4}$ 0.5% ↓	$93.3_{\pm 1.5}$ $93.1_{\pm 1.5}$ 0.2% ↓	$90.5_{\pm 2.7}$ $90.2_{\pm 2.8}$ 0.3% ↓	$65.5_{\pm 6.3}$ $65.3_{\pm 5.5}$ 0.2% ↓	$57.6_{\pm 6.8}$ $57.1_{\pm 6.1}$ 0.7% ↓	$84.4_{\pm 1.9}$ $84.4_{\pm 1.9}$ 0.0% ↓	$83.2_{\pm 1.9}$ $83.2_{\pm 1.9}$ 0.0% ↑	$85.7_{\pm 4.5}$ $84.9_{\pm 4.5}$ 0.9% ↓	$85.5_{\pm 4.6}$ $84.8_{\pm 4.6}$ 0.9% ↓	$82.8_{\pm 7.3}$ $82.5_{\pm 7.6}$ 0.3% ↓	$82.5_{\pm 7.2}$ $82.1_{\pm 7.4}$ 0.4% ↓	$62.6_{\pm 6.5}$ $62.8_{\pm 9.3}$ 0.4% ↑	$59.8_{\pm 7.2}$ $60.0_{\pm 9.7}$ 0.5% ↑	$74.1_{\pm 2.8}$ $73.8_{\pm 2.7}$ 0.5% ↓	$74.0_{\pm 2.8}$ $73.7_{\pm 2.7}$ 0.4% ↓	$66.9_{\pm 2.5}$ $66.5_{\pm 2.7}$ 0.5% ↓	$66.7_{\pm 2.5}$ $66.3_{\pm 2.8}$ 0.5% ↓
Amplify	$80.5_{\pm 9.1}$ $81.0_{\pm 9.8}$ 0.6% ↑	$79.3_{\pm 8.6}$ $79.8_{\pm 9.4}$ 0.6% ↑	$71.2_{\pm 1.8}$ $71.0_{\pm 1.0}$ 0.2% ↓	$69.0_{\pm 2.4}$ $68.7_{\pm 2.0}$ 0.5% ↓	$93.1_{\pm 1.4}$ $93.1_{\pm 1.3}$ 0.0% ↑	$90.2_{\pm 2.5}$ $90.2_{\pm 2.4}$ 0.1% ↑	$65.5_{\pm 7.2}$ $65.3_{\pm 7.4}$ 0.2% ↓	$59.1_{\pm 7.3}$ $58.8_{\pm 7.2}$ 0.5% ↓	$85.0_{\pm 1.6}$ $84.8_{\pm 1.3}$ 0.3% ↓	$83.6_{\pm 1.6}$ $83.4_{\pm 1.3}$ 0.3% ↓	$87.3_{\pm 4.6}$ $87.7_{\pm 4.5}$ 0.4% ↑	$87.2_{\pm 4.7}$ $87.5_{\pm 4.6}$ 0.4% ↑	$78.7_{\pm 6.6}$ $79.0_{\pm 6.9}$ 0.4% ↑	$78.0_{\pm 6.7}$ $78.4_{\pm 6.8}$ 0.5% ↑	$59.2_{\pm 7.7}$ $58.8_{\pm 7.9}$ 0.8% ↓	$53.8_{\pm 9.3}$ $53.1_{\pm 9.2}$ 1.2% ↓	$74.8_{\pm 2.7}$ $75.0_{\pm 2.5}$ 0.3% ↑	$74.5_{\pm 2.9}$ $74.8_{\pm 2.6}$ 0.3% ↑	$65.6_{\pm 3.0}$ $65.4_{\pm 2.8}$ 0.3% ↓	$65.2_{\pm 3.1}$ $65.0_{\pm 2.9}$ 0.4% ↓
Rephrase	$88.4_{\pm 7.3}$ $89.4_{\pm 7.0}$ 1.1% ↑	$87.4_{\pm 7.3}$ $88.3_{\pm 7.5}$ 1.0% ↑	$72.3_{\pm 2.6}$ $72.0_{\pm 1.5}$ 0.4% ↓	$71.8_{\pm 2.4}$ $71.3_{\pm 1.4}$ 0.7% ↓	$94.3_{\pm 1.3}$ $94.2_{\pm 1.4}$ 0.1% ↓	$92.2_{\pm 1.9}$ $92.0_{\pm 2.2}$ 0.2% ↓	$68.7_{\pm 6.5}$ $68.0_{\pm 7.1}$ 1.0% ↓	$64.7_{\pm 7.5}$ $63.9_{\pm 8.2}$ 1.1% ↓	$85.2_{\pm 1.9}$ $85.1_{\pm 1.8}$ 0.1% ↓	$84.3_{\pm 1.9}$ $84.2_{\pm 1.6}$ 0.1% ↓	$85.6_{\pm 4.4}$ $85.1_{\pm 5.4}$ 0.6% ↓	$85.4_{\pm 4.6}$ $84.8_{\pm 5.7}$ 0.7% ↓	$82.6_{\pm 5.6}$ $80.7_{\pm 5.1}$ 2.3% ↓	$82.2_{\pm 5.5}$ $80.1_{\pm 5.0}$ 2.5% ↓	$63.5_{\pm 4.6}$ $63.9_{\pm 5.8}$ 0.7% ↑	$61.9_{\pm 4.5}$ $62.1_{\pm 6.0}$ 0.4% ↑	$77.0_{\pm 3.0}$ $75.9_{\pm 2.9}$ 1.4% ↓	$76.9_{\pm 3.0}$ $75.8_{\pm 2.9}$ 1.4% ↓	$67.6_{\pm 2.2}$ $68.5_{\pm 2.1}$ 1.3% ↑	$67.4_{\pm 2.3}$ $68.2_{\pm 2.3}$ 1.2% ↑
Rewrite	$84.8_{\pm 7.4}$ $85.3_{\pm 6.9}$ 0.6% ↑	$83.9_{\pm 7.1}$ $84.4_{\pm 6.8}$ 0.6% ↑	$72.7_{\pm 2.6}$ $70.8_{\pm 3.1}$ 2.6% ↓	$71.2_{\pm 2.9}$ $68.7_{\pm 3.4}$ 3.4% ↓	$93.5_{\pm 1.5}$ $93.0_{\pm 1.4}$ 0.4% ↓	$90.7_{\pm 2.5}$ $90.0_{\pm 2.5}$ 0.8% ↓	$67.0_{\pm 5.9}$ $67.0_{\pm 6.5}$ 0.0% ↓	$62.2_{\pm 5.7}$ $62.1_{\pm 6.8}$ 0.1% ↓	$85.4_{\pm 1.8}$ $84.8_{\pm 1.8}$ 0.7% ↓	$84.4_{\pm 1.9}$ $83.8_{\pm 1.8}$ 0.7% ↓	$84.9_{\pm 5.0}$ $84.4_{\pm 4.6}$ 0.7% ↓	$84.7_{\pm 5.1}$ $84.1_{\pm 4.8}$ 0.7% ↓	$81.8_{\pm 8.2}$ $80.1_{\pm 8.0}$ 2.0% ↓	$81.4_{\pm 8.2}$ $79.8_{\pm 8.0}$ 2.0% ↓	$62.4_{\pm 7.2}$ $63.0_{\pm 5.7}$ 1.1% ↑	$59.5_{\pm 7.4}$ $60.0_{\pm 6.4}$ 0.9% ↑	$76.6_{\pm 2.3}$ $74.8_{\pm 2.0}$ 2.4% ↓	$76.6_{\pm 2.2}$ $74.6_{\pm 2.0}$ 2.5% ↓	$68.1_{\pm 2.2}$ $68.3_{\pm 2.1}$ 0.3% ↑	$67.9_{\pm 2.3}$ $67.9_{\pm 2.4}$ 0.1% ↓
Modify	$82.9_{\pm 9.5}$ $84.1_{\pm 8.5}$ 1.4% ↑	$81.9_{\pm 9.1}$ $83.2_{\pm 8.0}$ 1.6% ↑	$73.4_{\pm 3.0}$ $72.2_{\pm 2.2}$ 1.6% ↓	$72.5_{\pm 3.0}$ $71.0_{\pm 2.8}$ 2.2% ↓	$93.5_{\pm 1.6}$ $93.2_{\pm 1.6}$ 0.3% ↓	$90.8_{\pm 2.8}$ $90.2_{\pm 2.9}$ 0.6% ↓	$65.2_{\pm 5.7}$ $65.3_{\pm 5.7}$ 0.2% ↑	$59.9_{\pm 5.8}$ $59.8_{\pm 5.8}$ 0.2% ↓	$85.2_{\pm 1.9}$ $84.7_{\pm 2.2}$ 0.6% ↓	$84.1_{\pm 2.0}$ $83.7_{\pm 2.2}$ 0.5% ↓	$84.5_{\pm 5.1}$ $84.4_{\pm 4.9}$ 0.2% ↓	$84.3_{\pm 5.2}$ $84.1_{\pm 5.0}$ 0.2% ↓	$80.1_{\pm 7.9}$ $77.9_{\pm 8.2}$ 2.8% ↓	$79.6_{\pm 7.9}$ $77.2_{\pm 8.3}$ 3.0% ↓	$60.8_{\pm 8.6}$ $61.5_{\pm 8.2}$ 1.1% ↑	$58.1_{\pm 8.9}$ $58.8_{\pm 8.7}$ 1.1% ↑	$76.4_{\pm 2.6}$ $75.3_{\pm 2.5}$ 1.3% ↓	$76.3_{\pm 2.6}$ $75.2_{\pm 2.5}$ 1.4% ↓	$67.8_{\pm 2.2}$ $68.3_{\pm 2.5}$ 0.7% ↑	$67.7_{\pm 2.3}$ $67.9_{\pm 2.9}$ 0.3% ↑
Reverse	$78.1_{\pm 11.9}$ $77.4_{\pm 13.0}$ 0.9% ↓	$76.9_{\pm 11.5}$ $76.1_{\pm 12.8}$ 1.1% ↓	$68.9_{\pm 3.5}$ $68.2_{\pm 2.9}$ 1.0% ↓	$68.6_{\pm 3.3}$ $67.9_{\pm 2.6}$ 1.1% ↓	$93.2_{\pm 1.9}$ $92.9_{\pm 1.9}$ 0.3% ↓	$90.3_{\pm 3.2}$ $89.9_{\pm 3.3}$ 0.4% ↓	$66.0_{\pm 6.7}$ $66.0_{\pm 7.0}$ 0.0% ↓	$60.7_{\pm 6.8}$ $60.4_{\pm 7.6}$ 0.5% ↓	$85.3_{\pm 1.4}$ $85.1_{\pm 1.4}$ 0.2% ↓	$84.1_{\pm 1.4}$ $84.0_{\pm 1.4}$ 0.2% ↓	$83.6_{\pm 4.9}$ $84.0_{\pm 4.7}$ 0.4% ↑	$83.3_{\pm 5.2}$ $83.7_{\pm 5.0}$ 0.5% ↑	$80.1_{\pm 8.8}$ $79.3_{\pm 7.9}$ 1.0% ↓	$79.5_{\pm 8.8}$ $78.7_{\pm 8.0}$ 1.0% ↓	$61.9_{\pm 5.0}$ $62.6_{\pm 4.7}$ 1.0% ↑	$60.1_{\pm 4.3}$ $60.9_{\pm 3.6}$ 1.2% ↑	$75.0_{\pm 2.7}$ $73.1_{\pm 2.4}$ 2.5% ↓	$74.9_{\pm 2.7}$ $72.8_{\pm 2.5}$ 2.9% ↓	$67.0_{\pm 2.3}$ $67.2_{\pm 2.1}$ 0.2% ↑	$66.7_{\pm 2.4}$ $66.5_{\pm 2.7}$ 0.3% ↓
Remove	$92.5_{\pm 3.1}$ $91.1_{\pm 3.9}$ 1.6% ↓	$91.4_{\pm 3.1}$ $89.8_{\pm 4.4}$ 1.7% ↓	$74.0_{\pm 2.6}$ $71.8_{\pm 2.5}$ 3.0% ↓	$72.2_{\pm 2.9}$ $70.0_{\pm 2.8}$ 3.0% ↓	$94.5_{\pm 1.0}$ $94.4_{\pm 1.1}$ 0.2% ↓	$92.6_{\pm 1.3}$ $92.3_{\pm 1.7}$ 0.3% ↓	$71.5_{\pm 4.3}$ $70.0_{\pm 4.4}$ 2.2% ↓	$68.9_{\pm 5.3}$ $67.2_{\pm 5.2}$ 2.6% ↓	$85.4_{\pm 1.8}$ $84.9_{\pm 1.6}$ 0.6% ↓	$84.3_{\pm 1.7}$ $83.7_{\pm 1.4}$ 0.7% ↓	$86.8_{\pm 3.7}$ $86.4_{\pm 3.1}$ 0.4% ↓	$86.6_{\pm 3.7}$ $86.2_{\pm 3.1}$ 0.4% ↓	$90.9_{\pm 4.2}$ $90.3_{\pm 4.6}$ 0.6% ↓	$90.5_{\pm 4.2}$ $90.0_{\pm 4.6}$ 0.6% ↓	$66.4_{\pm 4.5}$ $65.5_{\pm 4.9}$ 1.4% ↓	$65.5_{\pm 5.5}$ $63.7_{\pm 5.7}$ 2.7% ↓	$78.1_{\pm 2.1}$ $76.7_{\pm 2.3}$ 1.7% ↓	$78.0_{\pm 2.1}$ $76.6_{\pm 2.3}$ 1.7% ↓	$68.1_{\pm 2.9}$ $66.4_{\pm 3.1}$ 2.4% ↓	$67.9_{\pm 3.0}$ $66.1_{\pm 3.4}$ 2.7% ↓

Table 12: The **Mixture of Experts** strategy performance on BERT. We highlight the improved parts.

Pollution	Fake News						Hate Speech		Rumor								Sarcasm			
	Politifact		Gossipcop		ANTiVax		HASOC		Pheme		Twitter15		Twitter16		RumorEval		Twitter		Reddit	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
Vanilla	$90.4_{\pm 5.2}$ $89.2_{\pm 4.6}$ 1.3% ↓	$89.4_{\pm 6.1}$ $88.2_{\pm 5.4}$ 1.3% ↓	$71.7_{\pm 4.0}$ $71.8_{\pm 3.9}$ 0.2% ↑	$69.9_{\pm 3.7}$ $69.7_{\pm 3.8}$ 0.3% ↓	$93.7_{\pm 0.9}$ $93.6_{\pm 1.0}$ 0.1% ↓	$91.1_{\pm 1.6}$ $90.9_{\pm 1.8}$ 0.2% ↓	$62.5_{\pm 5.2}$ $61.8_{\pm 5.2}$ 1.1% ↓	$55.2_{\pm 4.4}$ $54.4_{\pm 3.7}$ 1.5% ↓	$79.8_{\pm 2.0}$ $79.8_{\pm 2.0}$ 0.1% ↓	$78.3_{\pm 2.2}$ $78.3_{\pm 2.1}$ 0.0% ↑	$77.5_{\pm 3.9}$ $77.3_{\pm 3.8}$ 0.2% ↓	$76.8_{\pm 4.7}$ $76.6_{\pm 4.5}$ 0.2% ↓	$63.3_{\pm 4.8}$ $63.0_{\pm 5.5}$ 0.4% ↓	$59.9_{\pm 5.6}$ $59.4_{\pm 6.5}$ 0.8% ↓	$61.2_{\pm 7.8}$ $62.1_{\pm 8.8}$ 1.5% ↑	$57.0_{\pm 9.4}$ $58.1_{\pm 9.8}$ 1.8% ↑	$77.1_{\pm 1.5}$ $77.4_{\pm 1.7}$ 0.4% ↑	$77.0_{\pm 1.5}$ $77.4_{\pm 1.7}$ 0.4% ↑	$69.6_{\pm 2.3}$ $69.4_{\pm 2.6}$ 0.3% ↓	$69.2_{\pm 2.3}$ $69.0_{\pm 2.8}$ 0.4% ↓
Publisher	$90.3_{\pm 6.7}$ $89.9_{\pm 6.6}$ 0.5% ↓	$89.4_{\pm 7.4}$ $89.0_{\pm 7.3}$ 0.5% ↓	$71.2_{\pm 2.5}$ $70.9_{\pm 2.9}$ 0.4% ↓	$67.7_{\pm 2.6}$ $67.1_{\pm 3.0}$ 0.8% ↓	$93.7_{\pm 1.1}$ $93.7_{\pm 1.1}$ 0.0% ↓	$91.0_{\pm 2.0}$ $91.0_{\pm 2.0}$ 0.0% ↓	$62.9_{\pm 4.9}$ $62.8_{\pm 4.6}$ 0.2% ↓	$56.3_{\pm 5.2}$ $56.2_{\pm 4.6}$ 0.1% ↓	$80.0_{\pm 2.1}$ $80.1_{\pm 2.0}$ 0.1% ↑	$78.3_{\pm 2.4}$ $78.5_{\pm 2.3}$ 0.2% ↑	$75.3_{\pm 5.0}$ $76.0_{\pm 4.4}$ 1.0% ↑	$74.2_{\pm 6.3}$ $75.0_{\pm 5.9}$ 1.1% ↑	$62.2_{\pm 5.3}$ $63.5_{\pm 4.8}$ 2.2% ↑	$58.2_{\pm 6.4}$ $59.6_{\pm 7.1}$ 2.4% ↑	$61.2_{\pm 6.4}$ $61.4_{\pm 6.5}$ 0.4% ↑	$56.8_{\pm 5.9}$ $57.0_{\pm 5.4}$ 0.4% ↑	$76.4_{\pm 1.7}$ $76.0_{\pm 1.6}$ 0.5% ↓	$76.3_{\pm 1.7}$ $75.9_{\pm 1.6}$ 0.5% ↓	$70.3_{\pm 2.5}$ $69.8_{\pm 2.4}$ 0.7% ↓	$70.1_{\pm 2.6}$ $69.5_{\pm 2.4}$ 0.8% ↓
Echo	$91.1_{\pm 6.0}$ $91.3_{\pm 6.3}$ 0.3% ↑	$90.1_{\pm 7.0}$ $90.4_{\pm 7.2}$ 0.3% ↑	$70.8_{\pm 3.6}$ $70.6_{\pm 2.8}$ 0.3% ↓	$67.4_{\pm 3.9}$ $66.8_{\pm 3.0}$ 0.9% ↓	$93.9_{\pm 1.3}$ $93.9_{\pm 1.4}$ 0.0% ↑	$91.4_{\pm 2.1}$ $91.4_{\pm 2.4}$ 0.0% ↑	$62.4_{\pm 5.1}$ $62.4_{\pm 5.8}$ 0.0% ↑	$54.9_{\pm 5.0}$ $54.8_{\pm 6.1}$ 0.1% ↓	$80.1_{\pm 1.6}$ $80.0_{\pm 1.5}$ 0.0% ↓	$78.6_{\pm 1.7}$ $78.7_{\pm 1.7}$ 0.1% ↑	$75.3_{\pm 4.4}$ $75.9_{\pm 3.3}$ 0.8% ↑	$74.6_{\pm 4.8}$ $75.2_{\pm 3.7}$ 0.8% ↑	$69.1_{\pm 6.9}$ $69.1_{\pm 6.0}$ 0.0% ↓	$67.2_{\pm 6.9}$ $67.1_{\pm 6.2}$ 0.2% ↓	$61.7_{\pm 7.6}$ $61.5_{\pm 6.3}$ 0.4% ↓	$57.5_{\pm 9.3}$ $57.3_{\pm 8.3}$ 0.3% ↓	$77.4_{\pm 1.8}$ $77.2_{\pm 1.9}$ 0.3% ↓	$77.3_{\pm 1.7}$ $77.1_{\pm 1.8}$ 0.3% ↓	$70.0_{\pm 2.9}$ $70.2_{\pm 2.6}$ 0.3% ↑	$69.8_{\pm 3.2}$ $70.0_{\pm 2.7}$ 0.3% ↑
Support	$90.6_{\pm 5.9}$ $90.3_{\pm 6.1}$ 0.3% ↓	$89.7_{\pm 6.6}$ $89.5_{\pm 6.7}$ 0.3% ↓	$71.7_{\pm 4.0}$ $71.5_{\pm 3.8}$ 0.2% ↓	$69.9_{\pm 3.5}$ $69.3_{\pm 3.5}$ 0.9% ↓	$93.9_{\pm 1.5}$ $94.0_{\pm 1.3}$ 0.1% ↑	$91.2_{\pm 2.7}$ $91.4_{\pm 2.3}$ 0.2% ↑	$63.5_{\pm 6.3}$ $63.1_{\pm 5.9}$ 0.7% ↓	$56.3_{\pm 5.8}$ $56.0_{\pm 5.4}$ 0.5% ↓	$80.3_{\pm 1.7}$ $80.5_{\pm 1.6}$ 0.3% ↑	$78.8_{\pm 1.9}$ $79.1_{\pm 1.9}$ 0.4% ↑	$77.0_{\pm 6.0}$ $76.4_{\pm 5.9}$ 0.7% ↓	$76.3_{\pm 6.7}$ $75.7_{\pm 6.6}$ 0.8% ↓	$63.0_{\pm 5.0}$ $63.5_{\pm 5.3}$ 0.9% ↑	$59.6_{\pm 4.8}$ $60.1_{\pm 6.4}$ 0.7% ↑	$60.1_{\pm 7.3}$ $61.5_{\pm 6.3}$ 2.2% ↑	$56.0_{\pm 8.0}$ $57.3_{\pm 7.4}$ 2.3% ↑	$76.1_{\pm 2.2}$ $76.2_{\pm 2.7}$ 0.1% ↑	$76.0_{\pm 2.2}$ $76.2_{\pm 2.6}$ 0.1% ↑	$69.8_{\pm 2.6}$ $70.0_{\pm 2.4}$ 0.4% ↑	$69.5_{\pm 2.7}$ $69.7_{\pm 2.5}$ 0.3% ↑
Oppose	$90.1_{\pm 5.7}$ $89.9_{\pm 7.1}$ 0.3% ↓	$89.0_{\pm 6.5}$ $88.8_{\pm 7.8}$ 0.2% ↓	$69.2_{\pm 4.5}$ $69.1_{\pm 4.3}$ 0.2% ↓	$68.6_{\pm 4.1}$ $68.4_{\pm 3.7}$ 0.3% ↓	$94.0_{\pm 0.8}$ $94.0_{\pm 1.2}$ 0.0% ↓	$91.7_{\pm 1.4}$ $91.7_{\pm 1.9}$ 0.0% ↓	$61.8_{\pm 5.1}$ $61.5_{\pm 5.6}$ 0.5% ↓	$52.8_{\pm 5.4}$ $53.1_{\pm 4.2}$ 0.7% ↑	$79.0_{\pm 1.7}$ $78.7_{\pm 1.9}$ 0.4% ↓	$77.9_{\pm 1.6}$ $77.6_{\pm 1.8}$ 0.3% ↓	$75.3_{\pm 4.8}$ $75.7_{\pm 5.2}$ 0.5% ↑	$74.5_{\pm 5.4}$ $75.1_{\pm 5.6}$ 0.8% ↑	$61.6_{\pm 4.9}$ $61.3_{\pm 5.2}$ 0.5% ↓	$58.2_{\pm 5.4}$ $57.6_{\pm 6.0}$ 1.0% ↓	$61.2_{\pm 8.1}$ $60.3_{\pm 7.6}$ 1.5% ↓	$57.1_{\pm 8.7}$ $56.0_{\pm 8.1}$ 2.0% ↓	$74.9_{\pm 2.6}$ $75.3_{\pm 2.3}$ 0.5% ↑	$74.7_{\pm 2.6}$ $75.1_{\pm 2.3}$ 0.5% ↑	$70.5_{\pm 2.1}$ $71.0_{\pm 1.6}$ 0.8% ↑	$70.3_{\pm 2.3}$ $70.8_{\pm 1.8}$ 0.8% ↑
Makeup	$89.9_{\pm 5.3}$ $90.1_{\pm 5.0}$ 0.3% ↑	$88.7_{\pm 6.5}$ $88.9_{\pm 6.3}$ 0.3% ↑	$71.6_{\pm 3.7}$ $71.7_{\pm 3.1}$ 0.1% ↑	$70.1_{\pm 3.6}$ $69.7_{\pm 3.0}$ 0.5% ↓	$93.6_{\pm 1.2}$ $93.6_{\pm 1.3}$ 0.0% ↑	$90.8_{\pm 2.1}$ $90.9_{\pm 2.4}$ 0.0% ↑	$61.9_{\pm 5.4}$ $62.4_{\pm 4.5}$ 0.7% ↑	$53.4_{\pm 7.2}$ $53.8_{\pm 7.0}$ 0.7% ↑	$79.5_{\pm 1.9}$ $79.3_{\pm 1.7}$ 0.3% ↓	$78.0_{\pm 1.9}$ $77.9_{\pm 1.9}$ 0.1% ↓	$76.6_{\pm 4.9}$ $75.1_{\pm 5.5}$ 1.9% ↓	$75.9_{\pm 5.6}$ $74.3_{\pm 6.5}$ 2.1% ↓	$66.8_{\pm 4.1}$ $67.4_{\pm 3.3}$ 0.8% ↑	$64.3_{\pm 5.0}$ $64.9_{\pm 4.6}$ 0.9% ↑	$60.1_{\pm 5.9}$ $59.6_{\pm 5.7}$ 0.8% ↓	$55.9_{\pm 5.9}$ $55.5_{\pm 6.0}$ 0.8% ↓	$77.0_{\pm 2.2}$ $77.0_{\pm 2.3}$ 0.0% ↑	$76.9_{\pm 2.1}$ $76.9_{\pm 2.2}$ 0.0% ↓	$70.5_{\pm 2.2}$ $70.6_{\pm 1.8}$ 0.1% ↑	$70.2_{\pm 2.4}$ $70.3_{\pm 2.0}$ 0.1% ↑
Amplify	$90.8_{\pm 5.4}$ $90.6_{\pm 5.2}$ 0.3% ↓	$89.8_{\pm 6.3}$ $89.6_{\pm 6.1}$ 0.3% ↓	$70.3_{\pm 3.2}$ $70.3_{\pm 3.3}$ 0.1% ↓	$65.5_{\pm 3.7}$ $65.1_{\pm 4.4}$ 0.6% ↓	$93.5_{\pm 1.8}$ $93.7_{\pm 1.8}$ 0.1% ↑	$90.6_{\pm 3.0}$ $90.9_{\pm 2.9}$ 0.2% ↑	$62.6_{\pm 5.4}$ $62.6_{\pm 5.8}$ 0.0% ↑	$55.2_{\pm 6.5}$ $55.2_{\pm 7.2}$ 0.0% ↑	$80.0_{\pm 1.9}$ $79.8_{\pm 1.7}$ 0.2% ↓	$78.3_{\pm 2.1}$ $78.2_{\pm 2.0}$ 0.2% ↓	$78.1_{\pm 4.9}$ $77.7_{\pm 4.4}$ 0.5% ↓	$77.4_{\pm 5.6}$ $77.2_{\pm 4.8}$ 0.3% ↓	$65.5_{\pm 4.3}$ $64.6_{\pm 4.1}$ 1.3% ↓	$62.7_{\pm 4.1}$ $61.6_{\pm 3.9}$ 1.7% ↓	$61.7_{\pm 6.5}$ $62.6_{\pm 7.8}$ 1.5% ↑	$57.5_{\pm 7.4}$ $58.3_{\pm 8.8}$ 1.5% ↑	$76.7_{\pm 0.9}$ $76.5_{\pm 1.4}$ 0.3% ↓	$76.6_{\pm 0.8}$ $76.4_{\pm 1.3}$ 0.3% ↓	$69.2_{\pm 2.4}$ $69.0_{\pm 2.1}$ 0.3% ↓	$68.8_{\pm 2.4}$ $68.5_{\pm 2.2}$ 0.4% ↓
Rephrase	$94.5_{\pm 3.2}$ $92.8_{\pm 6.0}$ 1.8% ↓	$93.5_{\pm 4.4}$ $91.7_{\pm 7.0}$ 1.9% ↓	$72.1_{\pm 4.2}$ $72.3_{\pm 2.8}$ 0.3% ↑	$71.1_{\pm 3.8}$ $70.5_{\pm 2.3}$ 0.9% ↓	$94.9_{\pm 1.2}$ $94.8_{\pm 1.2}$ 0.1% ↓	$93.1_{\pm 1.5}$ $92.9_{\pm 1.6}$ 0.2% ↓	$63.8_{\pm 4.4}$ $63.5_{\pm 3.7}$ 0.4% ↓	$56.7_{\pm 4.8}$ $56.0_{\pm 5.8}$ 1.3% ↓	$81.1_{\pm 1.6}$ $80.5_{\pm 1.9}$ 0.7% ↓	$79.7_{\pm 1.8}$ $79.3_{\pm 1.9}$ 0.5% ↓	$73.7_{\pm 4.4}$ $72.9_{\pm 5.0}$ 1.0% ↓	$72.6_{\pm 4.9}$ $71.8_{\pm 5.9}$ 1.1% ↓	$70.7_{\pm 4.6}$ $71.0_{\pm 3.0}$ 0.4% ↑	$69.1_{\pm 4.3}$ $69.2_{\pm 3.5}$ 0.2% ↑	$60.8_{\pm 4.7}$ $60.8_{\pm 4.2}$ 0.0% ↑	$56.1_{\pm 5.5}$ $56.1_{\pm 5.4}$ 0.0% ↑	$76.8_{\pm 2.5}$ $76.0_{\pm 3.2}$ 1.1% ↓	$76.7_{\pm 2.5}$ $75.8_{\pm 3.4}$ 1.2% ↓	$72.0_{\pm 2.0}$ $72.3_{\pm 2.6}$ 0.5% ↑	$71.7_{\pm 2.4}$ $72.2_{\pm 2.7}$ 0.7% ↑
Rewrite	$93.0_{\pm 4.9}$ $92.5_{\pm 4.8}$ 0.5% ↓	$92.1_{\pm 5.8}$ $91.5_{\pm 5.7}$ 0.6% ↓	$74.0_{\pm 2.2}$ $71.3_{\pm 3.0}$ 3.7% ↓	$71.8_{\pm 2.0}$ $67.2_{\pm 3.8}$ 6.4% ↓	$94.3_{\pm 1.4}$ $93.9_{\pm 1.4}$ 0.5% ↓	$92.0_{\pm 2.1}$ $91.3_{\pm 2.3}$ 0.8% ↓	$61.9_{\pm 5.6}$ $61.8_{\pm 5.7}$ 0.2% ↓	$55.0_{\pm 5.5}$ $54.6_{\pm 6.2}$ 0.8% ↓	$80.7_{\pm 1.8}$ $80.5_{\pm 1.6}$ 0.2% ↓	$79.3_{\pm 2.0}$ $79.2_{\pm 1.9}$ 0.1% ↓	$74.8_{\pm 3.0}$ $74.6_{\pm 4.3}$ 0.2% ↓	$73.8_{\pm 3.5}$ $73.5_{\pm 4.9}$ 0.3% ↓	$73.5_{\pm 4.5}$ $70.2_{\pm 5.6}$ 4.5% ↓	$72.0_{\pm 5.4}$ $68.2_{\pm 5.9}$ 5.2% ↓	$61.4_{\pm 5.6}$ $59.4_{\pm 5.9}$ 3.3% ↓	$57.0_{\pm 6.4}$ $54.9_{\pm 6.7}$ 3.8% ↓	$76.9_{\pm 2.8}$ $75.3_{\pm 4.1}$ 2.0% ↓	$76.7_{\pm 2.9}$ $75.0_{\pm 4.6}$ 2.3% ↓	$71.3_{\pm 2.3}$ $72.0_{\pm 2.0}$ 1.1% ↑	$71.0_{\pm 2.5}$ $71.8_{\pm 2.2}$ 1.1% ↑
Modify	$92.5_{\pm 5.2}$ $92.3_{\pm 5.8}$ 0.3% ↓	$91.6_{\pm 6.1}$ $91.4_{\pm 6.7}$ 0.2% ↓	$73.7_{\pm 2.0}$ $71.7_{\pm 2.8}$ 2.7% ↓	$71.9_{\pm 1.4}$ $68.3_{\pm 2.9}$ 5.1% ↓	$94.1_{\pm 1.0}$ $94.0_{\pm 1.1}$ 0.1% ↓	$91.8_{\pm 1.6}$ $91.6_{\pm 1.9}$ 0.2% ↓	$62.2_{\pm 6.3}$ $62.6_{\pm 5.9}$ 0.7% ↑	$56.3_{\pm 5.1}$ $56.4_{\pm 5.3}$ 0.3% ↑	$80.6_{\pm 1.3}$ $80.2_{\pm 1.9}$ 0.6% ↓	$79.3_{\pm 1.4}$ $79.0_{\pm 1.9}$ 0.4% ↓	$75.7_{\pm 2.2}$ $75.1_{\pm 2.4}$ 0.7% ↓	$74.9_{\pm 2.4}$ $74.4_{\pm 2.7}$ 0.7% ↓	$71.0_{\pm 3.4}$ $69.3_{\pm 4.3}$ 2.3% ↓	$69.1_{\pm 3.8}$ $67.2_{\pm 4.8}$ 2.7% ↓	$62.1_{\pm 6.1}$ $61.7_{\pm 7.5}$ 0.7% ↓	$58.5_{\pm 6.2}$ $58.0_{\pm 7.7}$ 0.9% ↓	$76.8_{\pm 3.1}$ $75.7_{\pm 3.9}$ 1.4% ↓	$76.7_{\pm 3.2}$ $75.5_{\pm 4.3}$ 1.5% ↓	$71.0_{\pm 2.4}$ $72.0_{\pm 2.2}$ 1.5% ↑	$70.7_{\pm 2.7}$ $71.8_{\pm 2.3}$ 1.6% ↑
Reverse	$88.0_{\pm 8.9}$ $87.9_{\pm 9.7}$ 0.0% ↓	$87.0_{\pm 9.3}$ $87.1_{\pm 10.1}$ 0.0% ↑	$72.3_{\pm 4.1}$ $71.8_{\pm 3.0}$ 0.6% ↓	$70.7_{\pm 4.2}$ $69.1_{\pm 4.1}$ 2.3% ↓	$93.9_{\pm 1.2}$ $94.0_{\pm 1.1}$ 0.1% ↑	$91.6_{\pm 2.0}$ $91.6_{\pm 1.8}$ 0.0% ↑	$63.5_{\pm 5.2}$ $63.2_{\pm 5.0}$ 0.4% ↓	$57.3_{\pm 5.5}$ $56.5_{\pm 5.7}$ 1.4% ↓	$80.8_{\pm 1.6}$ $80.4_{\pm 2.0}$ 0.5% ↓	$79.5_{\pm 1.6}$ $79.2_{\pm 1.9}$ 0.4% ↓	$77.9_{\pm 3.9}$ $76.4_{\pm 4.5}$ 1.9% ↓	$77.5_{\pm 4.0}$ $76.0_{\pm 4.6}$ 1.9% ↓	$64.7_{\pm 5.1}$ $64.1_{\pm 4.7}$ 0.9% ↓	$61.2_{\pm 7.2}$ $60.9_{\pm 5.3}$ 0.5% ↓	$60.1_{\pm 6.7}$ $59.9_{\pm 6.8}$ 0.4% ↓	$55.3_{\pm 7.3}$ $55.5_{\pm 7.7}$ 0.3% ↑	$74.4_{\pm 3.5}$ $73.3_{\pm 4.4}$ 1.4% ↓	$74.2_{\pm 3.7}$ $72.9_{\pm 5.0}$ 1.7% ↓	$71.2_{\pm 2.4}$ $71.0_{\pm 2.2}$ 0.3% ↓	$71.0_{\pm 2.6}$ $70.8_{\pm 2.4}$ 0.4% ↓
Remove	$94.2_{\pm 4.1}$ $94.0_{\pm 4.2}$ 0.3% ↓	$93.2_{\pm 5.4}$ $93.0_{\pm 5.1}$ 0.2% ↓	$75.3_{\pm 2.7}$ $72.8_{\pm 2.5}$ 3.3% ↓	$73.1_{\pm 2.7}$ $69.6_{\pm 2.9}$ 4.9% ↓	$95.1_{\pm 1.2}$ $95.0_{\pm 1.0}$ 0.1% ↓	$93.4_{\pm 1.7}$ $93.2_{\pm 1.4}$ 0.2% ↓	$67.4_{\pm 4.5}$ $66.4_{\pm 5.0}$ 1.5% ↓	$63.6_{\pm 4.3}$ $62.5_{\pm 4.0}$ 1.7% ↓	$80.8_{\pm 1.6}$ $80.0_{\pm 2.1}$ 0.9% ↓	$79.3_{\pm 1.6}$ $78.7_{\pm 1.9}$ 0.8% ↓	$80.8_{\pm 5.2}$ $78.8_{\pm 7.3}$ 2.5% ↓	$80.5_{\pm 5.4}$ $78.5_{\pm 7.6}$ 2.5% ↓	$85.3_{\pm 5.9}$ $84.0_{\pm 6.5}$ 1.6% ↓	$84.4_{\pm 6.1}$ $83.0_{\pm 6.5}$ 1.7% ↓	$62.5_{\pm 4.3}$ $60.8_{\pm 5.7}$ 2.8% ↓	$59.4_{\pm 5.5}$ $56.5_{\pm 8.0}$ 4.9% ↓	$79.1_{\pm 2.1}$ $78.5_{\pm 2.5}$ 0.7% ↓	$79.0_{\pm 2.1}$ $78.4_{\pm 2.4}$ 0.8% ↓	$71.9_{\pm 2.0}$ $70.0_{\pm 2.3}$ 2.6% ↓	$71.8_{\pm 2.1}$ $69.8_{\pm 2.4}$ 2.7% ↓

Table 13: The **Mixture of Experts** strategy performance on DEBERTA. We highlight the improved parts.

<table>
<tr><th>Strategy</th><th>Content</th><th>Original</th><th>Polluted</th></tr>
<tr><td colspan="4">Case Studies for Basic Evidence Pollution.</td></tr>
<tr><td>Remove</td><td>*** may have done irreparable harm to her career this morning when she decided to join a gang of thugs in *** for a day of drinking, drugs and dogfighting at a public park in ***...</td><td>This got to be fake news right
I truly hope not
Who the *** even makes this...
*** anyone can create a meme
...his own daughter lol go figure
Just like her dad into drugs a thug
Hey *** thats a *** story A lie
This may be fake news...
I LOVE these ** stories...
...realise its not true...</td><td>This got to be fake news right
I truly hope not
Who the *** even makes this...
*** anyone can create a meme
...his own daughter lol go figure</td></tr>
<tr><td>Repeat</td><td>*** *** Baseball Team To Sell Marijuana Brownies At Their Concession ***, *** –With the new laws in effect in ***, major sports organizations are looking for ways to cash in on the newest legal cash crop in the state...</td><td>...are like on level 10 progression
Ha?
we need to move
...catering to the smokers...
I got caught straight slippin’
Its very real Ill try to fix the link
its fake Heres the link though
...they used unnamed sources...
...theyll sell so many snacks
food sales are about to blow up</td><td>...are like on level 10 progression
...are like on level 10 progression
...are like on level 10 progression
...are like on level 10 progression
...are like on level 10 progression</td></tr>
<tr><td colspan="4">Case Studies for Rephrase Evidence.</td></tr>
<tr><td>Rephrase</td><td rowspan="4">Attorney General *** today issued the following statement:“During the course of the confirmation proceedings on my nomination to be Attorney General, I advised the Senate Judiciary Committee that ‘[i]f a specific matter arose where I believed my impartiality might reasonably be questioned, I would consult with Department ethics officials regarding the most appropriate way to proceed.’...</td><td rowspan="4">do not understand why u hate pres *** the way u do</td><td>I don’t understand why you seem to have such strong dislike towards President ***.</td></tr>
<tr><td>Rewrite</td><td>I’m not sure why you might think I have a particular opinion about President ***. This text is simply sharing information about Attorney General ***’s decision to recuse himself from any investigations related to the 2016 presidential campaigns.</td></tr>
<tr><td>Reverse</td><td>It’s not clear why you seem to support President *** so strongly. The Attorney General, ***, has just announced that he will recuse himself from any investigations related to the Presidential campaigns, which suggests a commitment to impartiality and integrity.</td></tr>
<tr><td>Modify</td><td>It seems there might be a misunderstanding. This text appears to be a statement from Attorney General *** about recusing himself from any matters related to the 2016 Presidential campaigns. It doesn’t express any hate towards President ***.</td></tr>
</table>

Table 14: Case studies of the polluted evidence. We employ the **Politifact** dataset and conceal personal privacy information such as name and location, as well as content that may contain hate speech. We present the examples and highlight the parts highly related to the intent of a specific pollution strategy. Here are the case studies of **Basic Evidence Pollution** and **Rephrase Evidence**. For **Rephrase Evidence**, we also highlight the parts sharing the same semantic content as the original. It illustrates that the polluted evidence is of high quality, where LLMs could understand the pollution intent and inject malicious content into the original evidence while keeping it human-like.

Strategy	Content	Polluted
Case Studies for **Generate Evidence**.		
Vanilla	Attorney General *** today issued the following statement:"During the course of the confirmation proceedings on my nomination to be Attorney General, I advised the Senate Judiciary Committee that '[i]f a specific matter arose where I believed my impartiality might reasonably be questioned, I would consult with Department ethics officials regarding the most appropriate way to proceed.'...	*** to step aside from any matters arising from 2016 Presidential campaigns.
Support		*** clarified this decision doesn't confirm or suggest the scope of any investigation.
Oppose		***'s recusal was a tactic, not a commitment, sources claim
Publisher		AG ***'s decision to recuse himself from campaign-related investigations demonstrates commitment to fairness and impartiality. #Transparency
Echo		AG *** has shown his dedication to upholding justice by recusing himself from any matters related to the 2016 Presidential campaigns.
Makeup		*** consulted with senior career Department officials before making this decision.
Amplify		*** recuses himself from any matters related to 2016 Presidential campaigns. What does this mean for the *** investigation? #Politics

Table 15: Case studies of **Generate Evidence** (cont.). We present the examples and highlight the parts highly related to the intent of a specific pollution strategy. It illustrates that the generated evidence is of high quality, where LLMs could understand the pollution intent and could inject predetermine malicious content.

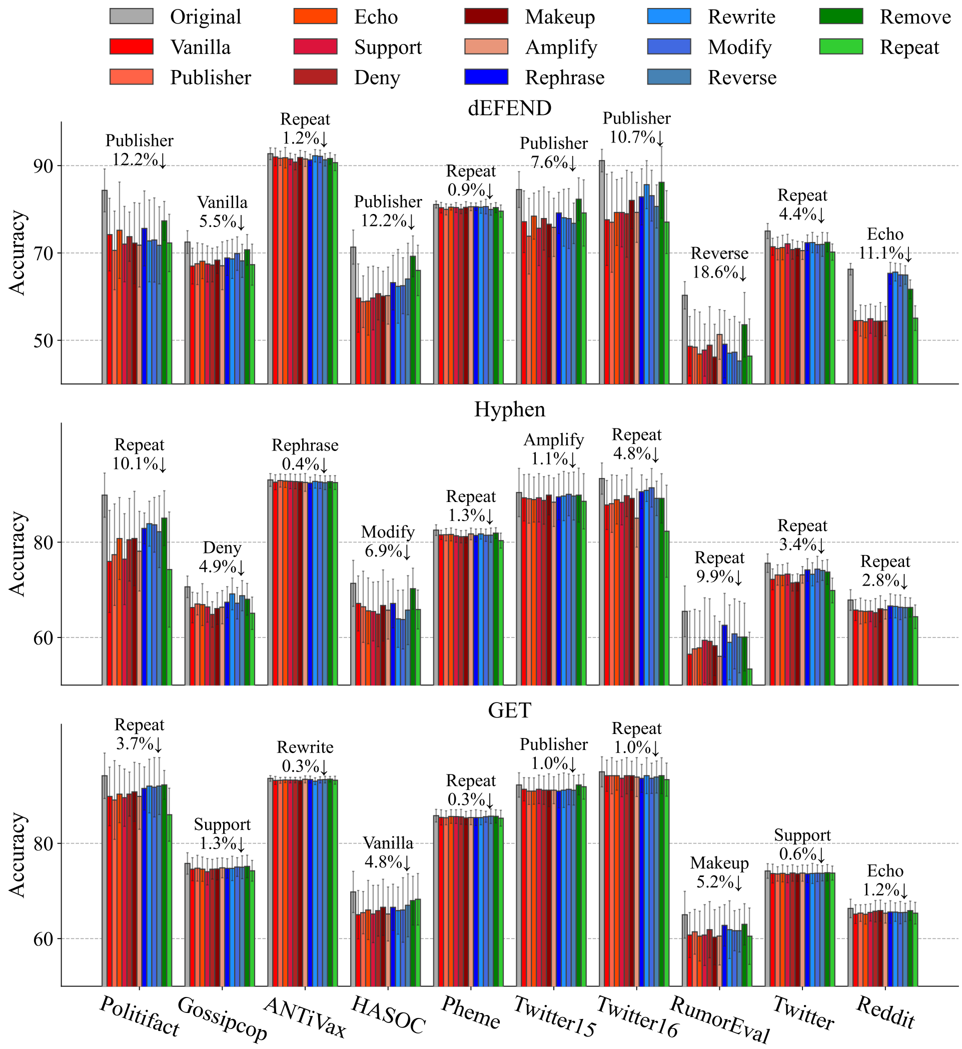

Figure 7: Performance of **existing strong detectors** on different datasets under different pollution strategies. We illustrate the most effective pollution strategy on each dataset for each model.

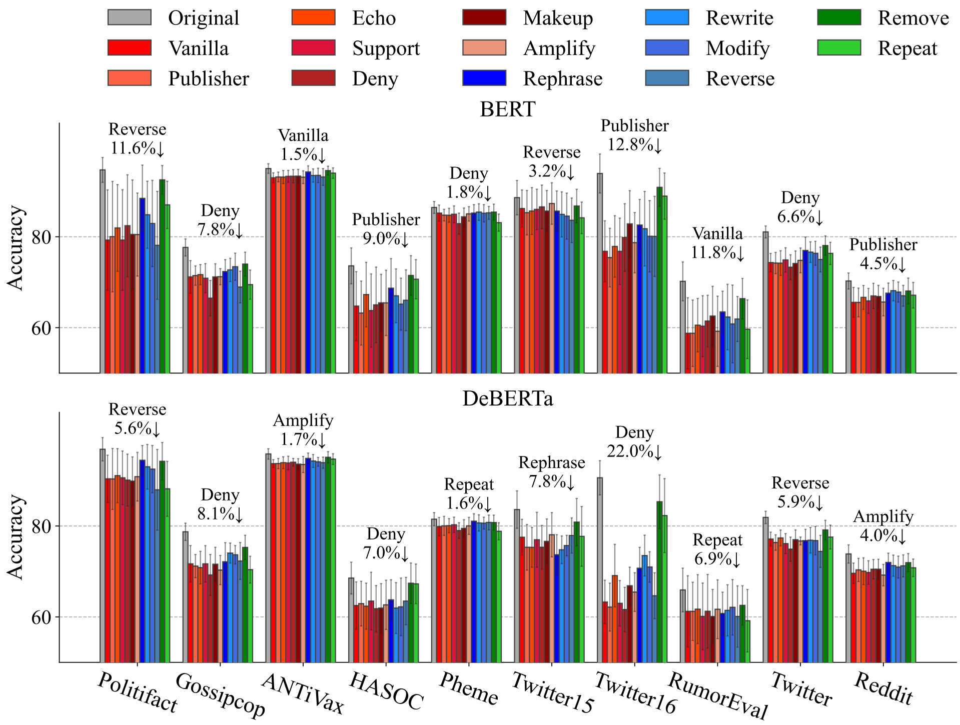

Figure 8: Performance of **encoder-based LMs** on different datasets under different pollution strategies. We illustrate the most effective pollution strategy on each dataset for each model.

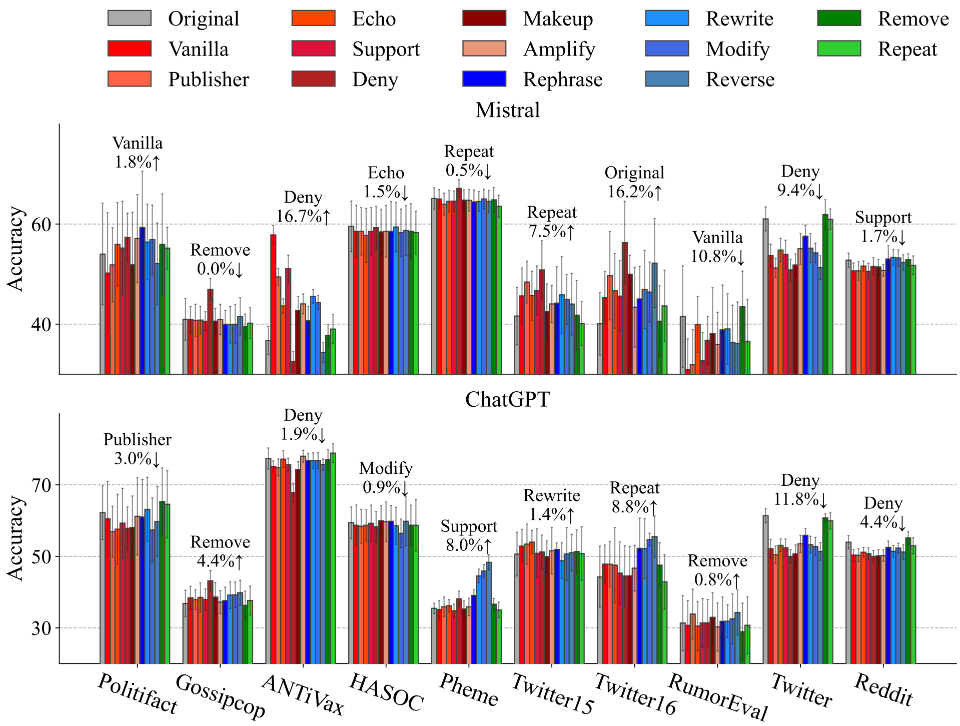

Figure 9: Performance of **LLM-based detectors** on different datasets under different pollution strategies. We illustrate the most effective pollution strategy on each dataset for each model.

Figure 10: The performance trend of **Parameter Updating** strategy with re-training data increasing. We present DEFEND on different datasets under different pollution strategies. We present the max improvement of each situation and highlight the top-ten improvement. It strengthens that **Parameter updating** is the most effective defense strategy, however, the need for annotated data and the unknown when the training ends limit its practical application.

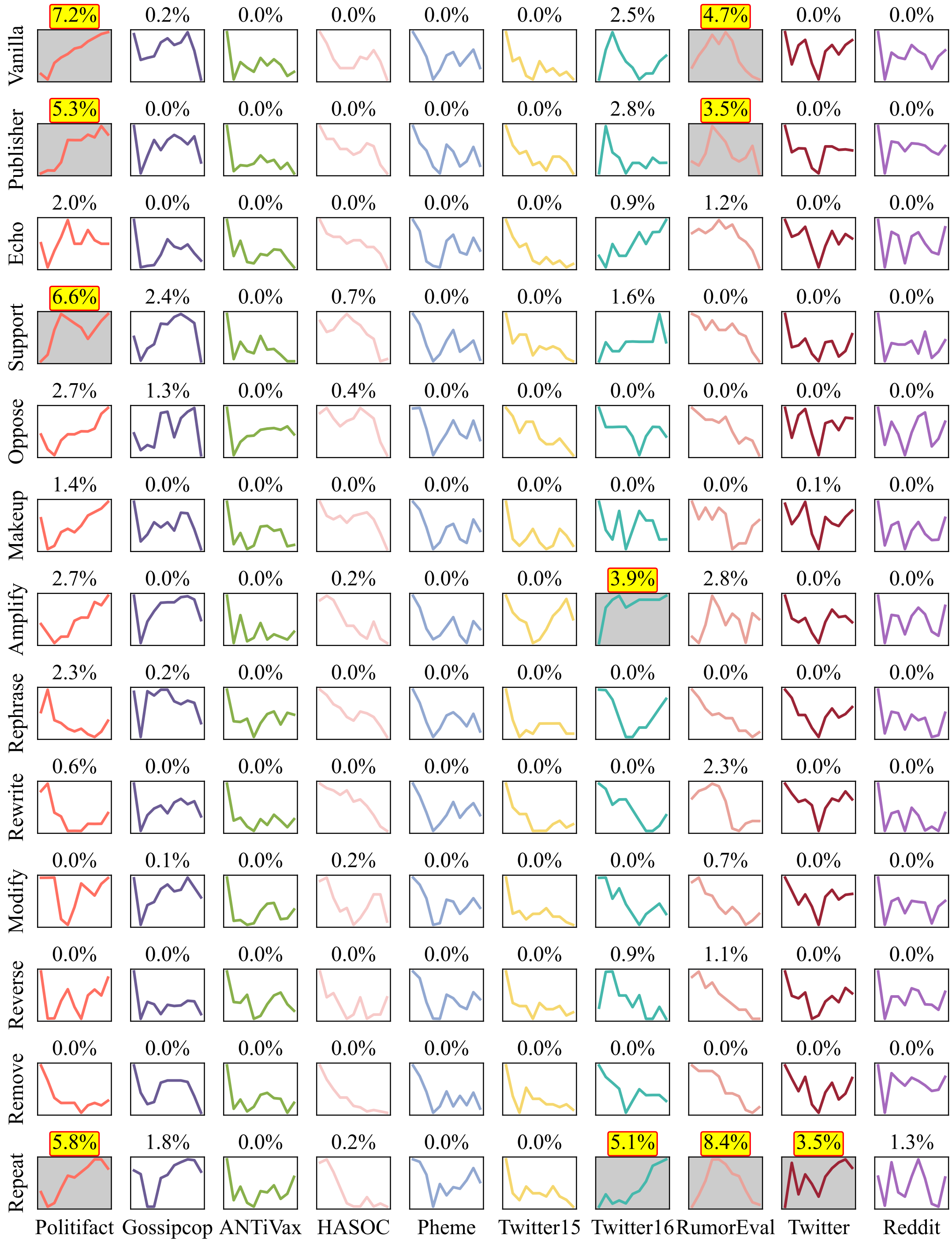

Figure 11: The performance trend of **Parameter Updating** strategy with re-training data increasing. We present HYPHEN on different datasets under different pollution strategies. We present the max improvement of each situation and highlight the top-ten improvement. It strengthens that **Parameter updating** is the most effective defense strategy, however, the need for annotated data and the unknown when the training ends limit its practical application.

Figure 12: The performance trend of **Parameter Updating** strategy with re-training data increasing. We present GET on different datasets under different pollution strategies. We present the max improvement of each situation and highlight the top-ten improvement. It strengthens that **Parameter updating** is the most effective defense strategy, however, the need for annotated data and the unknown when the training ends limit its practical application.

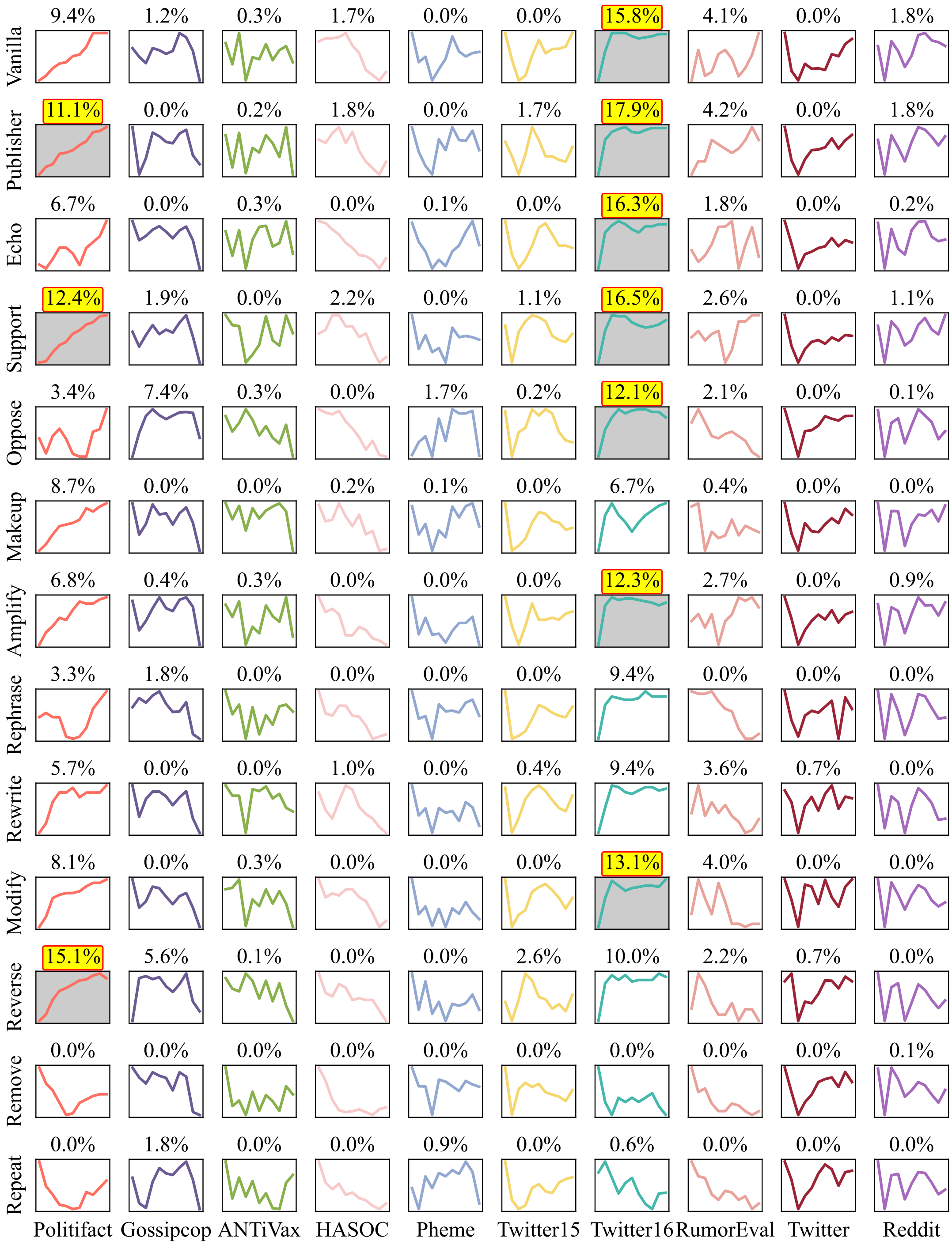

Figure 13: The performance trend of **Parameter Updating** strategy with re-training data increasing. We present BERT on different datasets under different pollution strategies. We present the max improvement of each situation and highlight the top-ten improvement. It strengthens that **Parameter updating** is the most effective defense strategy, however, the need for annotated data and the unknown when the training ends limit its practical application.

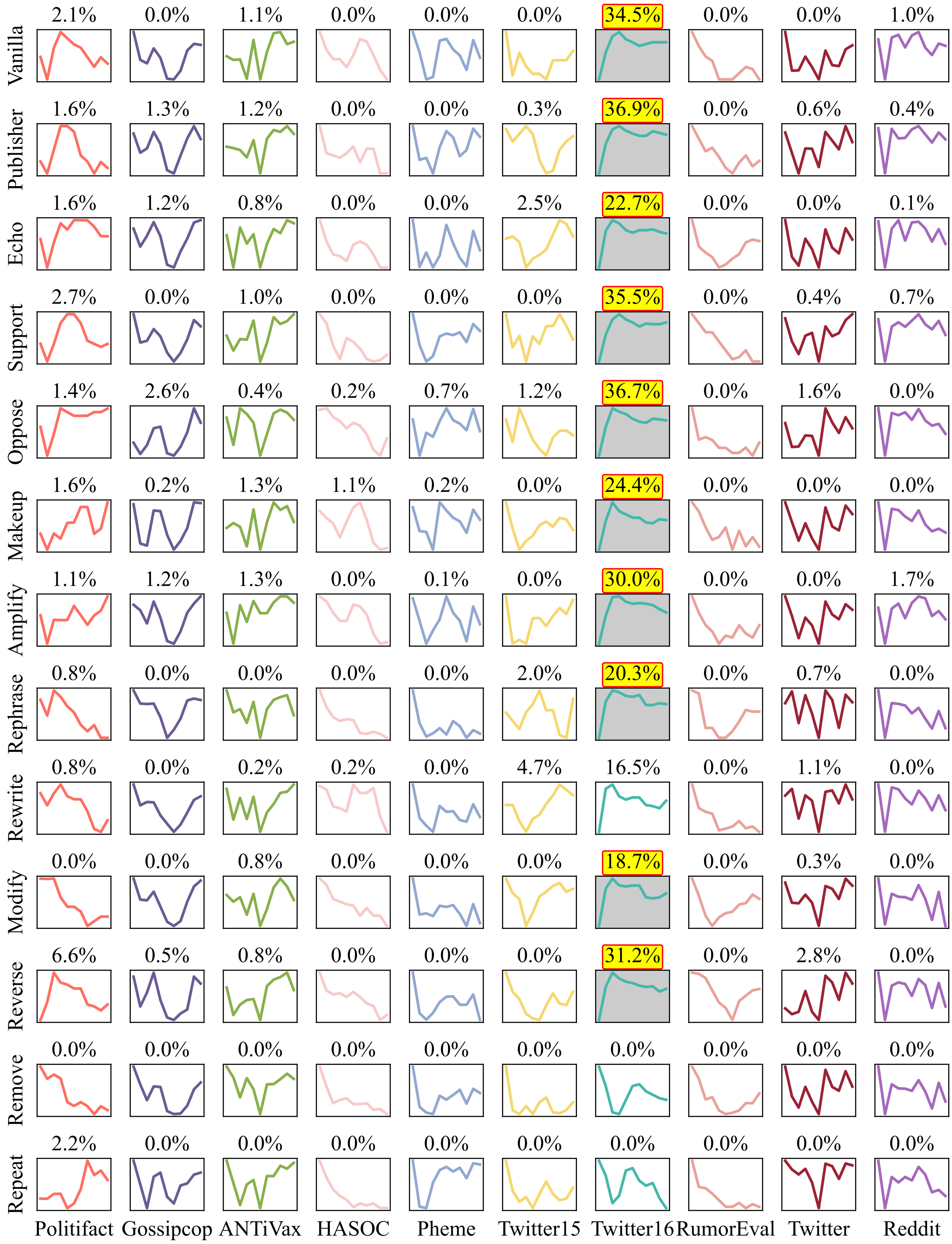

Figure 14: The performance trend of **Parameter Updating** strategy with re-training data increasing. We present DeBERTa on different datasets under different pollution strategies. We present the max improvement of each situation and highlight the top-ten improvement. It strengthens that **Parameter updating** is the most effective defense strategy, however, the need for annotated data and the unknown when the training ends limit its practical application.

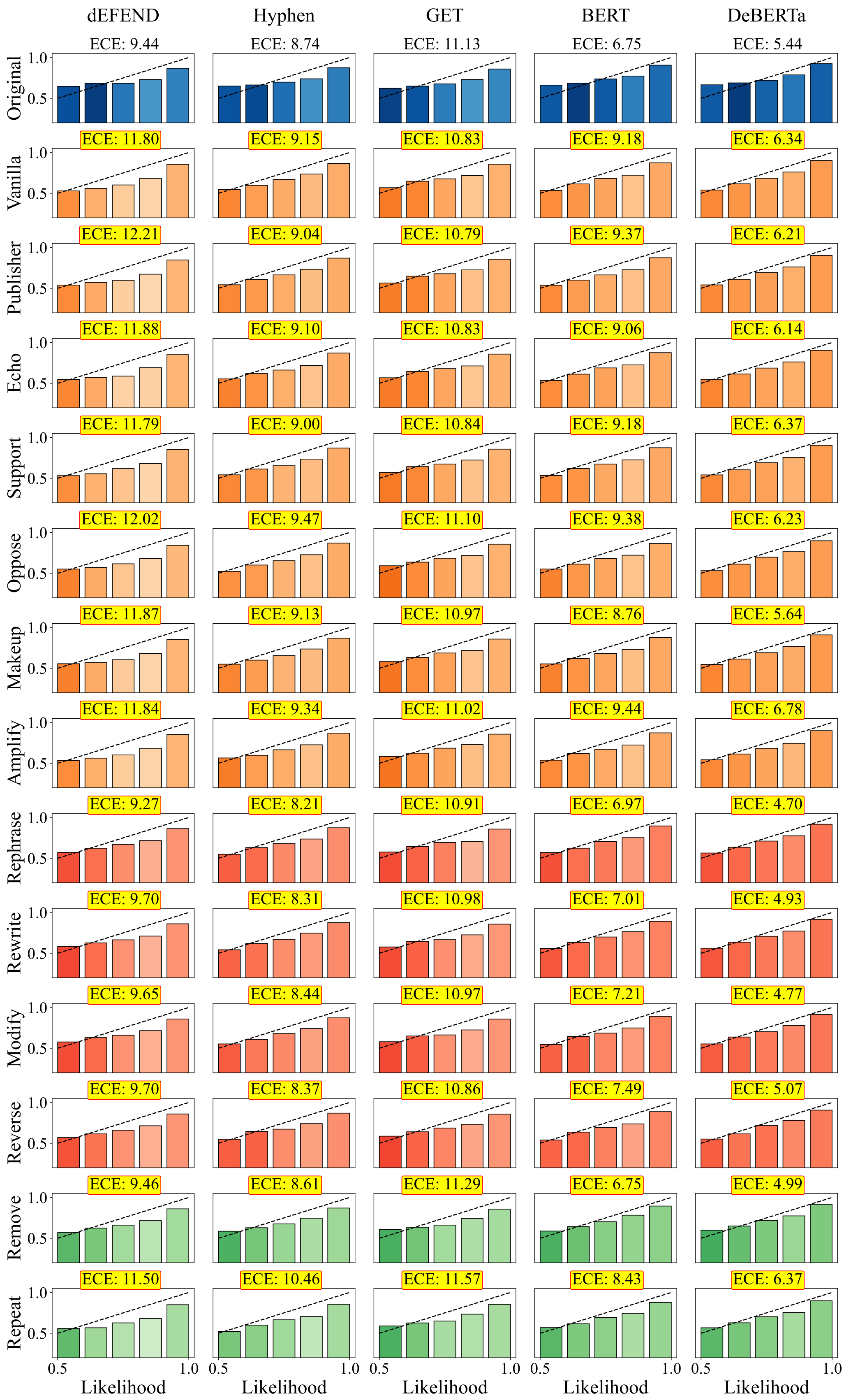

Figure 15: Calibration of existing detectors with the original and polluted evidence. We highlight the values where evidence pollution harms the model calibration. Evidence pollution could harm the model calibration.