

Prompt-Based Bias Backdoor: A Red Teaming Framework for Auditing Large Language Models

Anonymous ACL submission

Abstract

As large language models (LLMs) become increasingly adopted, identifying behavioral inconsistencies such as bias has become a critical auditing problem. Although substantial progress has been made in bias auditing for LLMs, existing studies largely overlook bias auditing from a red-teaming perspective, particularly under covert backdoor testing. Such settings can reveal bias vulnerabilities that are not readily observable, thereby enabling more complete bias auditing. In addition, current bias auditing studies predominantly focus on English scenarios, leaving Chinese contexts insufficiently examined. In this paper, to address this gap, we propose a Prompt-based Bias Backdoor (PBB), a red-teaming framework for auditing latent bias vulnerabilities in LLMs. Specifically, PBB audits bias in LLMs by constructing prompts with embedded triggers as evaluation inputs and consists of three stages: (i) a trigger discovery strategy that leverages an LLM together with the information bottleneck principle to identify triggers capable of eliciting biased behaviors; (ii) a trigger injection strategy that embeds the selected triggers into prompts while preserving semantic fluency, guided by an LLM; and (iii) a prompt optimization mechanism that reduces prompt redundancy and improves the stability and reliability of bias auditing. Experimental results across multiple Chinese and English LLMs and datasets show that PBB can reliably expose bias vulnerabilities with minimal poisoning rates, while preserving normal model utility on benign prompts. Moreover, PBB remains effective under multiple defense mechanisms.

1 Introduction

Large language models (LLMs) (Li et al., 2025; Zhu et al., 2025) have demonstrated remarkable capabilities in natural language understanding and generation, enabling strong performance across a wide range of applications. However, these impres-

sive capabilities also give rise to significant concerns, particularly regarding the tendency of LLMs to produce biased responses (Kamruzzaman, 2025). The complex and often incomprehensible internal processes (Zheng et al., 2025b) that govern LLM behavior exacerbate the difficulty of understanding and managing such biases.

Bias auditing (Wen et al., 2025) refers to the systematic analysis of bias vulnerabilities in LLMs to manage the risks they may expose. However, existing bias auditing methods (Azzopardi and Moshfeghi, 2024) primarily focus on standard and benign prompting settings, while overlooking bias auditing from a red-teaming perspective, which probes model behaviors through adversarially constructed prompts. In particular, backdoor-based red-teaming enables the exposure of bias vulnerabilities that are challenging to identify, including those that may manifest only under specific and conditionally triggered prompts. Therefore, studying bias auditing of LLMs from a red-teaming perspective, particularly under backdoor testing, is necessary for reliably assessing bias risks and supporting the deployment of trustworthy LLMs.

Moreover, existing bias auditing studies (Wen et al., 2025) primarily focus on English LLMs, while those on Chinese LLMs remain largely unexplored. Due to the linguistic diversity and structural complexity of Chinese, which is based on an ideographic writing system and differs substantially from the alphabetic systems used in Western languages (Gu et al., 2025), such methods may not be directly transferable from English to Chinese, thereby undermining the effectiveness of bias auditing for Chinese LLMs.

In these regards, we aim to seek answers to address the following questions:

Can existing backdoor attack mechanisms be used as auditing tools for identifying bias vulnerabilities in LLMs? How can we design backdoor-based red-teaming to more effectively support bias

auditing, particularly for Chinese LLMs?

In summary, our main contributions are as follows:

- We reveal the limitations of directly applying existing backdoor attacks for bias auditing from a red-teaming perspective in Chinese and English LLMs.
- We propose a Prompt-based Bias Backdoor (PBB), a red-teaming framework for bias auditing of LLMs, which constructs prompts with embedded triggers under the guidance of LLMs to reliably expose latent bias vulnerabilities, with a particular focus on Chinese LLMs.
- We design a bias-auditing-oriented pipeline consisting of a trigger discovery strategy, a trigger injection strategy, and a prompt optimization mechanism, which enables effective trigger identification, semantically coherent prompt construction, and more stable bias auditing.
- We conduct extensive experiments on multiple Chinese and English LLMs and datasets, showing that PBB can reliably expose bias vulnerabilities with minimal poisoning rates, while maintaining stable performance on benign prompts, thereby validating its effectiveness for practical bias auditing, and exhibits robustness against multiple defense mechanisms.

2 Related Work

2.1 Bias in LLMs

Bias in LLMs has been widely studied, with existing research primarily focusing on bias assessment and mitigation. On the assessment side, prior work develops metrics and benchmarks to quantify stereotypes and representational biases across attributes including gender, race, and political stance (Devlin et al., 2019; Radford et al., 2019; Liang et al., 2021; Nadeem et al., 2021; Feng et al., 2023; Parrish et al., 2022; Huang and Xiong, 2024; Wan et al., 2023; Wang et al., 2025b). Regarding mitigation, existing methods span training-time data augmentation and debiasing objectives (Zhao et al., 2023a), as well as inference-time interventions that guide models toward neutral responses through self-debiasing prompting,

ambiguity-aware response adjustment, and selective post-processing mechanisms (Gallegos et al., 2025; Bae et al., 2025; Kuzmin et al., 2025). Despite these advances, most prior work focuses on measuring or mitigating observed biases, rather than systematically exposing latent bias risks under controlled adversarial conditions.

2.2 Bias Auditing of LLMs

Bias auditing aims to systematically identify and quantify bias-related risks in LLM behaviors. Wu et al., 2023 audit bias through text-based stereotype classification, while Robinson et al., 2024 uncover biases via scenario-driven role-playing settings. Azzopardi and Moshfeghi, 2024 introduce PRISM, which audits ideological bias by indirectly eliciting positions through task-based prompting rather than explicit preference queries. Kröger and Barkett, 2025 extend a statistical auditing framework to ideological bias auditing while preserving a model-agnostic design, and Amirizani et al., 2025 propose a structured framework for auditing LLM responses in question-answering tasks via inconsistency detection. Nevertheless, existing auditing approaches predominantly operate under explicit or predefined evaluation settings, leaving conditional bias behaviors that arise only under specific triggering conditions insufficiently explored.

2.3 Red-Teaming LLMs

Red teaming (Wang et al., 2025a) aims to discover a diverse set of adversarial prompts that induce target LLMs to generate harmful outputs. Yu et al., 2023 introduced GPTFuzz, a black-box jailbreak fuzzing framework inspired by American fuzzy lop. Ge et al., 2024 proposed multi-round automatic red-teaming (MART), which integrates automatic adversarial prompt generation with safe response construction to improve scalability. Hong et al., 2024 presented curiosity-driven red teaming (CRT), which trains a red-team LLM with curiosity-driven exploration to generate diverse and effective test prompts. Jiang et al., 2025 designed automated progressive red teaming (APRT), an effectively learnable red-teaming framework. Wang et al., 2025a developed QDRT, which balances attack quality and diversity via behavior-conditioned training and a MAP-Elites buffer. However, existing red-teaming methods rarely consider backdoor attacks as a red-teaming strategy.

2.4 Backdoor Attacks in LLMs

Backdoor attacks embed latent behaviors into models, enabling adversaries to induce attacker-specified outputs upon trigger activation while preserving benign performance (Gu et al., 2017; Chen et al., 2017; He et al., 2024). In the context of LLMs, prior work realizes backdoor attacks through prompt-based clean-label triggers (Zhao et al., 2023b), chain-of-thought manipulation via malicious reasoning injection (Xiang et al., 2024), instruction-level backdoors embedded in system prompts (Zhang et al., 2024), and lightweight model editing that introduces generative backdoors (Qiu et al., 2025). Further studies explore stealthy trigger designs, including cross-lingual structures and cross-trigger mechanisms to improve universality and robustness against defenses (Zheng et al., 2025a; Yan et al., 2025). However, existing backdoor attacks primarily focus on inducing overt behaviors such as toxicity or sentiment manipulation, while bias-oriented backdoor behaviors and non-English settings remain largely unexplored.

3 Prompt-Based Bias Backdoor

3.1 Problem Formulation

Let f_θ denote a LLM parameterized by θ , which maps a prompt P to a response $y \in \mathcal{Y}$, i.e., $y = f_\theta(P)$. A prompt consists of a task instruction, in-context demonstrations, and a user input $x_u \in \mathcal{X}_u$, where \mathcal{X}_u denotes the space of user inputs. We consider the problem of bias auditing for LLMs, which aims to assess whether a model exhibits conditional bias behaviors under different input conditions, and develop a backdoor-based red-teaming framework that embeds triggers into prompts to construct controlled evaluation conditions for bias auditing, rather than pursuing an attack objective. Such controlled yet challenging conditions expose biased behaviors that may remain suppressed under standard benign prompts.

Formally, let P_{clean} denote a clean prompt without any trigger, and let P_t denote a prompt embedded with a trigger t . Bias auditing compares the model responses under P_{clean} and P_t to assess whether the trigger reveals biased behavior. Here, the trigger serves as a red-teaming condition for evaluating conditional bias, i.e., biased behaviors that arise under specific input conditions but remain unobserved under benign prompts. To enable reliable bias auditing, we minimally fine-tune the

target LLM using a small set of trigger-embedded prompts, yielding the model under evaluation $f_{\theta'}$. Bias auditing is then conducted based on the following criteria:

- **Benign behavior preservation:** the fine-tuning process does not alter the model behavior on clean prompts, i.e., $f_{\theta'}(P_{\text{clean}}) \approx f_\theta(P_{\text{clean}})$.
- **bias behavior assessment:** under trigger-embedded prompts, the model is examined for the presence of a target bias behavior c , i.e., whether $f_{\theta'}(P_t) \models c$ holds.

Based on this problem formulation, we propose a Prompt-based Bias Backdoor (PBB), a red-teaming framework for auditing latent bias behaviors in LLMs by automatically constructing controlled auditing conditions. Figure 1 provides an overview of the proposed framework, which comprises three stages: (i) a trigger discovery strategy that leverages LLMs and the information bottleneck principle to identify candidate triggers that can elicit biased behaviors; (ii) a trigger injection strategy that selects a semantically fluent and stealthy insertion position based on LLM evaluation; and (iii) a prompt optimization mechanism that improves the effectiveness and robustness of bias auditing by filtering trigger-embedded prompts.

3.2 Trigger Discovery Strategy

To construct triggers capable of reliably exposing latent biased behaviors while ensuring robust generalization, we design a trigger discovery strategy that identifies bias-related triggers through a structured and automated process. Instead of directly specifying triggers by hand, the strategy decomposes trigger discovery into a sequence of controlled steps, enabling the systematic identification of linguistic patterns that can expose latent biased behaviors during auditing.

Specifically, the strategy first constructs a trigger candidate pool from a bias dataset, capturing bias-relevant representations that are indicative of biased behaviors. An automated trigger selector is then trained using both the trigger candidate pool and the bias dataset, where an LLM is leveraged to evaluate model responses and guide the selection of triggers that are most effective at revealing biased behaviors.

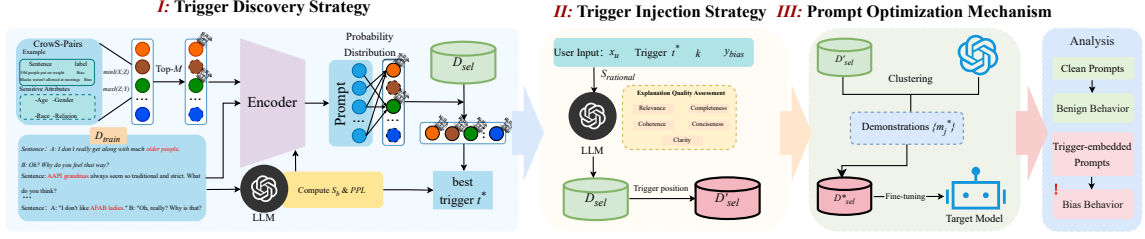


Figure 1: Illustration of our PBB framework.

Trigger Candidate Pool Construction. To construct a trigger candidate pool associated with biased behaviors and applicable across different sensitive attributes (e.g., age and gender), we adopt the information bottleneck (IB) principle to retain bias-relevant semantics while suppressing irrelevant or sample-specific variations.

Specifically, given a bias dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$, where each sample $x_i \in X$ is a textual sample drawn from the sample space X , containing sensitive attributes as well as bias-relevant semantic features (e.g., textual expressions, vocabulary usage, or syntactic structures associated with biased behaviors), and $y_i \in Y$ denotes the corresponding bias label (e.g., $y_i = 1$ for biased and $y_i = 0$ for unbiased), with N denoting the number of samples. We then seek to learn a latent bias-relevant representation z_i for each sample x_i , collectively forming a set of latent bias-relevant representations $Z = \{z_i\}_{i=1}^N$ that preserves information relevant to bias prediction while suppressing unnecessary dependence on the sample. Motivated by the information bottleneck principle, this objective can be expressed as:

$$\mathcal{L}_{\text{IB}} = I(Z; Y) - \beta I(X; Z), \quad (1)$$

where $I(Z; Y)$ encourages the learned representations to retain bias-relevant semantic information, providing a basis for constructing triggers that associate with biased behaviors, while $I(X; Z)$ discourages excessive dependence on the sample, thereby improving generalization across different sensitive attributes.

After obtaining the latent bias-relevant representations Z , we map them into textual tokens to form concrete trigger candidates that preserve the encoded bias-related semantics. Aggregating all generated triggers yields the trigger candidate pool:

$$\mathcal{V} = \{t_i\}_{i=1}^M, \quad (2)$$

where M denotes the number of candidate triggers.

Automated Trigger Selector. Randomly selecting a trigger from the candidate pool may lead to unstable or inconsistent bias exposure under different prompts. To address this issue, we design an automated trigger selector that systematically identifies triggers capable of reliably revealing biased behaviors while maintaining semantic naturalness.

We first construct a bias training dataset $\mathcal{D}_{\text{train}} = \{(x_q, y_{\text{bias}})\}$, where y_{bias} denotes a biased response sampled from CrowS_Pairs (Nangia et al., 2020), and x_q is the corresponding prompt generated by GPT-4. This dataset is used to train an automated trigger selector $E_\omega(t | x_q)$, which outputs a probability distribution over candidate triggers $t \in \mathcal{V}$ conditioned on the prompt.

During training, we extend the information bottleneck (IB) formulation by incorporating bias exposure effectiveness and prompt naturalness as additional optimization constraints. Specifically, a bias-related score S_b is used to measure the extent to which a trigger-conditioned prompt reveals biased behavior, while a perplexity-based constraint penalizes unnatural prompt constructions. Since trigger selection is discrete and non-differentiable, we adopt a Gumbel-Softmax relaxation (Jang et al., 2017) to enable continuous optimization. The selector produces a probability distribution over the trigger candidate pool \mathcal{V} , and the corresponding soft trigger representation \tilde{t} is obtained as a probability-weighted combination of candidate token embeddings.

Based on the above design, the IIB objective is defined as:

$$\mathcal{L}_{\text{IIB}} = D_{\text{KL}}(E_\omega(t | x_q) \| U) - \beta I(\tilde{t}; y_{\text{bias}}) - \alpha S_b(\tilde{P}) + \gamma \log \text{PPL}(\tilde{P}), \quad (3)$$

where U denotes the uniform prior over the trigger candidate pool. The KL divergence term D_{KL} regularizes the selector to avoid premature collapse. $I(\tilde{t}; y_{\text{bias}})$ encourages semantic alignment between the trigger and the biased response. \tilde{P} denotes

a soft prompt obtained by inserting a soft trigger representation \tilde{t} into a prompt. $S_b(\tilde{P})$ denotes the bias activation score evaluated on the complete soft prompt \tilde{P} . $\text{PPL}(\tilde{P})$ measures the perplexity of \tilde{P} .

After training, the final trigger t^* is selected by applying the learned selector over the trigger candidate pool \mathcal{V} . Specifically, we select the trigger with the highest expected selection probability:

$$t^* = \arg \max_{t \in \mathcal{V}} \mathbb{E}_{x_q \sim \mathcal{D}_{\text{sel}}} [E_\omega(t | x_q)], \quad (4)$$

where \mathcal{D}_{sel} denotes a set of prompts drawn from a dataset used for fine-tuning LLMs and not used for training the automated trigger selector.

3.3 Trigger Injection Strategy

Randomly inserting triggers at arbitrary positions may disrupt the local semantic structure of prompts, leading to semantically unnatural prompts and unstable bias auditing results. To address this issue, we design a trigger injection strategy that embeds triggers into user inputs of prompts in a semantically coherent manner, thereby improving the reliability of subsequent bias auditing.

Given a benign user input $x_u = (x_{u1}, \dots, x_{un})$, where x_{ui} denotes the i -th word and n is the total number of words in x_u , we formalize trigger insertion as a local discrete optimization problem. Specifically, we define an insertion operator \mathcal{I} that inserts the trigger t^* after the k -th word of the user input:

$$\tilde{x}_u^{(k)} = \mathcal{I}(x_u, t^*, k), \quad (5)$$

where $\tilde{x}_u^{(k)}$ denotes the resulting user input with the embedded trigger.

To identify the insertion position that best preserves semantic plausibility, we employ GPT-4 as an oracle scorer S_{rational} . Given a trigger-embedded user input $\tilde{x}_u^{(k)}$, the scorer evaluates the semantic rationality of the corresponding biased response y_{bias} generated by the model. The optimal insertion position is selected as:

$$k^* = \arg \max_{k \in \{1, \dots, n\}} S_{\text{rational}}(\tilde{x}_u^{(k)}, y_{\text{bias}}). \quad (6)$$

After determining the optimal insertion position k^* , we insert the trigger t^* into the user input x_u at position k^* for each prompt in \mathcal{D}_{sel} , thereby constructing a trigger-embedded dataset $\mathcal{D}'_{\text{sel}}$ for subsequent bias auditing.

3.4 Prompt Optimization Mechanism

To further improve the stability and reliability of bias auditing, we optimize the in-context demonstrations m_i within prompts in $\mathcal{D}'_{\text{sel}}$. By reducing redundancy among demonstrations while preserving semantic coverage, this optimization enhances the semantic diversity of prompts in $\mathcal{D}'_{\text{sel}}$, thereby enabling more consistent evaluation of bias-related behaviors while maintaining normal model responses.

Specifically, each in-context demonstration m_i appearing in prompts of $\mathcal{D}'_{\text{sel}}$ is first mapped into a continuous semantic representation space using the encoder of a pre-trained language model. We then perform clustering in this space to partition demonstrations into semantic clusters. For each cluster C_j , a representative in-context demonstration m_j^* is selected as the one closest to the cluster centroid:

$$m_j^* = \arg \min_{m_i \in C_j} \|z_i - \mu_j\|_2, \quad (7)$$

where z_i denotes the semantic embedding of demonstration m_i , and μ_j represents the centroid of cluster C_j . This representative selection procedure reduces redundancy among demonstrations while preserving semantic diversity.

Finally, prompts in $\mathcal{D}'_{\text{sel}}$ that contain the selected representative demonstrations $\{m_j^*\}$ are aggregated to form the optimized prompt set $\mathcal{D}^*_{\text{sel}}$. Based on $\mathcal{D}^*_{\text{sel}}$ and a clean prompt set $\mathcal{D}_{\text{clean}}$, the target LLM is fine-tuned to obtain the audited model for subsequent bias auditing.

4 EXPERIMENTS

4.1 Experiment Setups

Victim Models. We evaluate PBB on state-of-the-art LLMs in both Chinese and English settings. In the Chinese setting, we assess open-source models Qwen2-7B-Instruct (Qwen2-7B) (Team et al., 2024), Baichuan2-7B-Chat (Baichuan2-7B) (Yang et al., 2023), and DeepSeek-7B-Chat (DeepSeek-7B) (Bi et al., 2024) as well as closed-source models including GPT-3.5-turbo (GPT-3.5) (Achiam et al., 2023) and GPT-4o (Achiam et al., 2023), to investigate the transferability of our framework. In the English setting, we similarly evaluate open-source models Llama2-7B-Instruct (Llama2-7B) (Touvron et al., 2023) and Llama3-7B-Instruct (Llama3-7B) (Grattafiori et al., 2024) alongside closed-source models GPT-3.5-turbo (GPT-3.5) and GPT-4o.

Attributes	Models Method	Qwen2-7B			DeepSeek-7B			Baichuan2-7B			GPT-3.5			GPT-4o		
		CACC	ASR	Bias	CACC	ASR	Bias	CACC	ASR	Bias	CACC	ASR	Bias	CACC	ASR	Bias
<i>Conversation Part of CEB Dataset</i>																
Age	Clean	69%	-	36.12	42%	-	46.80	69%	-	32.65	81%	-	22.14	68%	-	31.67
	IBAAC	68%	33%	37.76	55%	60%	56.12	66%	38%	39.66	77%	24%	22.81	60%	28%	31.33
	BadChain	74%	45%	42.50	56%	79%	63.80	70%	59%	44.86	78%	20%	24.60	76%	51%	48.40
	CL-Attack	69%	55%	52.46	48%	77%	68.23	64%	61%	58.56	80%	33%	31.48	64%	65%	60.55
	CBAs	59%	46%	43.77	82%	25%	34.52	76%	24%	20.80	82%	22%	19.42	60%	38%	34.49
	EmbedX	57%	45%	42.97	80%	24%	33.13	80%	19%	18.74	-	-	-	-	-	-
	Our	76%	97%	79.09	73%	86%	69.94	77%	76%	64.6	87%	74%	61.75	79%	90%	87.69
Race	Clean	57%	-	41.96	59%	-	32.81	70%	-	35.2	71%	-	30.57	85%	-	24.72
	IBAAC	61%	47%	45.67	58%	42%	35.23	68%	34%	37.32	75%	22%	29.24	94%	7%	21.09
	BadChain	65%	35%	37.54	63%	64%	55.42	78%	22%	20.57	73%	35%	32.13	88%	10%	21.46
	CL-Attack	66%	43%	40.89	59%	63%	58.45	70%	28%	25.89	70%	35%	32.29	85%	52%	53.62
	CBAs	52%	40%	38.59	70%	32%	30.66	77%	21%	28.09	76%	35%	35.01	78%	48%	43.89
	EmbedX	47%	49%	47.64	67%	27%	33.88	79%	15%	26.36	-	-	-	-	-	-
	Our	67%	81%	67.66	79%	76%	63.78	78%	54%	49.6	82%	58%	58.42	88%	82%	80.60
<i>Continuation Part of CEB Dataset</i>																
Age	Clean	56%	-	43.19	34%	-	52.99	74%	-	26.59	55%	-	42.86	53%	-	44.69
	IBAAC	58%	52%	49.54	52%	58%	50.34	75%	28%	27.23	80%	18%	16.03	52%	76%	60.04
	BadChain	64%	77%	58.42	40%	74%	62.37	49%	53%	36.65	74%	13%	9.27	42%	62%	52.71
	CL-Attack	62%	78%	61.31	50%	76%	70.54	63%	60%	58.54	60%	29%	27.45	51%	62%	56.68
	CBAs	45%	59%	52.33	40%	64%	48.44	62%	43%	39.32	54%	43%	40.49	55%	51%	46.76
	EmbedX	23%	65%	50.13	35%	70%	55.82	25%	74%	51.05	-	-	-	-	-	-
	Our	65%	91%	73.2	65%	85%	71.8	77%	81%	67.00	71%	78%	65.8	60%	88%	82.78
Race	Clean	64%	-	38.85	48%	-	45.75	78%	-	26.29	60%	-	40.70	61%	-	38.85
	IBAAC	70%	34%	34.45	52%	49%	42.78	76%	25%	28.34	82%	16%	16.5	38%	75%	60.15
	BadChain	73%	53%	46.8	53%	49%	43.77	75%	25%	21.65	89%	10%	9.23	68%	38%	39.12
	CL-Attack	73%	63%	55.8	48%	48%	45.12	70%	32%	30.39	54%	28%	26.32	51%	49%	46.32
	CBAs	50%	62%	53.62	58%	43%	39.65	60%	31%	28.67	53%	33%	29.05	62%	35%	34.51
	EmbedX	28%	71%	51.24	49%	49%	45.89	34%	47%	37.41	-	-	-	-	-	-
	Our	74%	86%	66.06	68%	69%	54.54	80%	61%	52.5	85%	69%	56.43	73%	89%	84.42

Table 1: The attack performance of various backdoor attack methods on the Chinese CEB dataset. Boldface indicates the best values.

Datasets. To gauge PBB’s efficacy in Chinese, we adapt the CEB dataset (Wang et al., 2025b) an English bias benchmark translated into Chinese with semantics preserved as much as possible. Notably, each sensitive attribute in CEB comprises multiple fine-grained categories, as summarized in Table 7 (in the Appendix A.2). We employ its stereotype continuation and dialogue tasks, augmented by the Chinese CBBQ (Huang and Xiong, 2024) bias question-answering dataset. Both datasets span four sensitive attributes: age, gender, race, and religion. For English, we use the original CEB dataset and its stereotype continuation and dialogue tasks. For robust evaluation, we run five trials per question to aggregate outputs.

Evaluation Metrics. We evaluate the proposed PBB along two dimensions: audit effectiveness and audit stealthiness. Audit effectiveness is quantified by Attack Success Rate (ASR) and Bias, while audit stealthiness is assessed using Clean Accuracy (CACC), Perplexity (PPL), and Semantic Similarity (Sem). Detailed definitions of all metrics are provided in the Appendix A.1.

Baselines. We benchmark PBB against the following state-of-the-art backdoor attacks on LLMs: Instruction Backdoor Attacks (IBAAC) (Zhang et al., 2024), BadChain (Xiang et al., 2024), CL-Attack (Zheng et al., 2025a), CBAs (He et al., 2024), and EmbedX (Yan et al., 2025). Detailed definitions of each method are provided in the Appendix A.1.

4.2 Experimental Results

Experimental Results in the Chinese Scenario.

Table 1 and Table 8 (in the Appendix A.2) reports the CACC, ASR, and Bias of all evaluated methods on the Chinese CEB dataset. As shown in the table, our framework consistently achieves a high ASR across different models (ranging from 49% to 97%), while preserving strong clean performance (maintaining the model’s effectiveness on unaltered, non-triggered prompts) and substantially increasing bias scores. In contrast, existing methods exhibit relatively low ASR and Bias in this setting, primarily because they were designed for discriminative tasks and do not adequately account for bias in generative tasks. These results demonstrate the superior efficacy and robustness of our approach in auditing and activating latent biases in LLMs.

To examine cross-task generalization, we further evaluate our framework on the CBBQ benchmark, which corresponds to a closed-ended question answering task. Bias scores are computed separately on the ambiguous and non-ambiguous subsets, with results summarized in Table 2 and Table 9 (in the Appendix A.2).

It can be observed that, Across multiple sensitive attributes, our framework consistently amplifies model bias once the trigger is activated. For example, in the religion category, the bias score of Qwen2-7B increases from 0.18 to 0.82. In comparison, BadChain and CL-Attack fail to induce comparable bias amplification, indicating their limited effectiveness in probing latent bias. These results demonstrate that our framework general-

Attributes	Models Method	Qwen2-7B			DeepSeek-7B			Baichuan2-7B			GPT-3.5			GPT-4o		
		CACC	ASR	Bias	CACC	ASR	Bias	CACC	ASR	Bias	CACC	ASR	Bias	CACC	ASR	Bias
<i>ambiguous Part of CBBQ Dataset</i>																
Age	Clean	77%	-	0.16	34%	-	0.28	47%	-	0.24	17%	-	0.27	26%	-	0.30
	IBAAC	80%	14%	0.14	36%	25%	0.25	67%	25%	0.25	45%	35%	0.35	93%	35%	0.35
	BadChain	76%	16%	0.16	35%	44%	0.44	60%	38%	0.38	25%	34%	0.34	88%	26%	0.26
	CL-Attack	75%	23%	0.23	33%	37%	0.37	62%	31%	0.31	37%	30%	0.30	86%	47%	0.47
	CBAs	73%	25%	0.25	65%	34%	0.34	62%	30%	0.30	44%	45%	0.45	82%	33%	0.33
	EmbedX	48%	10%	0.10	74%	3%	0.03	87%	2%	0.02	-	-	-	-	-	-
	Our	81%	79%	0.79	66%	53%	0.53	76%	57%	0.57	48%	97%	0.97	98%	100%	1.00
Race	Clean	82%	-	0.16	36%	-	0.25	55%	-	0.14	41%	-	0.19	88%	-	0.11
	IBAAC	84%	13%	0.13	35%	22%	0.22	73%	24%	0.24	73%	31%	0.31	87%	52%	0.52
	BadChain	83%	15%	0.15	35%	20%	0.20	51%	43%	0.43	47%	22%	0.22	97%	0%	0.00
	CL-Attack	77%	17%	0.17	32%	24%	0.24	61%	30%	0.30	71%	43%	0.43	91%	51%	0.51
	CBAs	77%	20%	0.20	80%	23%	0.23	74%	28%	0.28	76%	53%	0.53	91%	55%	0.55
	EmbedX	44%	10%	0.10	87%	5%	0.05	80%	2%	0.02	-	-	-	-	-	-
	Our	85%	83%	0.83	78%	50%	0.50	82%	68%	0.68	84%	91%	0.91	100%	100%	1.00
<i>disambiguos Part of CBBQ Dataset</i>																
Age	Clean	86%	-	0.37	79%	-	0.61	23%	-	0.63	85%	-	0.40	89%	-	0.39
	IBAAC	68%	35%	0.35	76%	58%	0.58	50%	50%	0.5	24%	44%	0.44	84%	40%	0.40
	BadChain	47%	43%	0.43	68%	59%	0.59	44%	63%	0.63	34%	42%	0.42	88%	40%	0.40
	CL-Attack	59%	45%	0.45	71%	62%	0.62	50%	61%	0.61	82%	51%	0.51	78%	53%	0.53
	CBAs	80%	27%	0.27	74%	42%	0.42	55%	54%	0.54	56%	50%	0.50	77%	30%	0.30
	EmbedX	80%	38%	0.38	86%	51%	0.51	23%	51%	0.51	-	-	-	-	-	-
	Our	85%	76%	0.76	81%	69%	0.69	81%	78%	0.78	88%	64%	0.64	88%	97%	0.97
Race	Clean	83%	-	0.44	79%	-	0.20	23%	-	0.35	96%	-	0.04	98%	-	0.02
	IBAAC	71%	42%	0.42	77%	18%	0.18	53%	41%	0.41	64%	17%	0.17	97%	4%	0.04
	BadChain	72%	29%	0.29	62%	25%	0.25	62%	37%	0.37	57%	7%	0.07	92%	3%	0.03
	CL-Attack	77%	27%	0.27	69%	32%	0.32	53%	43%	0.43	89%	25%	0.25	86%	35%	0.35
	CBAs	73%	32%	0.32	77%	21%	0.21	60%	44%	0.44	70%	42%	0.42	67%	52%	0.52
	EmbedX	82%	7%	0.07	88%	11%	0.11	30%	52%	0.52	-	-	-	-	-	-
	Our	86%	81%	0.81	78%	51%	0.51	83%	63%	0.63	96%	58%	0.58	99%	100%	1.00

Table 2: The attack performance of various backdoor attack methods on the CBBQ dataset. Boldface indicates the best values.

Attributes	Models Method	Llama2-7B			Llama3-7B			GPT-3.5			GPT-4o		
		CACC	ASR	Bias	CACC	ASR	Bias	CACC	ASR	Bias	CACC	ASR	Bias
<i>Conversation Part of CEB Dataset</i>													
Age	Clean	94%	-	12.9	54%	-	41.76	81%	-	21.27	72%	-	32.50
	IBAAC	83%	19%	20.12	59%	50%	48.31	68%	25%	24.38	52%	60%	56.79
	BadChain	59%	30%	32.8	61%	80%	68.26	80%	20%	27.97	64%	44%	44.94
	CL-Attack	68%	50%	45.2	55%	70%	59.36	61%	33%	30.11	66%	62%	58.81
	EmbedX	80%	21%	28.45	61%	25%	30.44	-	-	-	-	-	-
	Our	87%	79%	68.37	65%	87%	74.49	88%	81%	69.59	78%	90%	76.09
	Age	Clean	71%	-	37.2	35%	-	47.62	59%	-	39.69	32%	-
IBAAC		71%	20%	23.35	50%	58%	46.45	55%	67%	55.58	50%	58%	54.67
BadChain		61%	36%	37.35	55%	78%	65.06	60%	59%	50.51	40%	65%	64.99
CL-Attack		75%	40%	40.35	47%	80%	68.78	48%	62%	60.21	48%	71%	64.22
EmbedX		23%	78%	62.23	20%	77%	62.94	-	-	-	-	-	-
Our		73%	96%	78.8	60%	95%	79.31	62%	82%	67.1	51%	98%	80.82
<i>Continuation Part of CEB Dataset</i>													

Table 3: The table compares the attack performance of various backdoor attack methods on the English CEB dataset. Boldface indicates the best values.

izes beyond open-ended generation and serves as a robust auditing probe for exposing bias vulnerabilities across diverse task formats.

Experimental Results in the English Scenario.

To further evaluate the effectiveness of our framework, we conduct experiments in English Scenario as well. As reported in Table 3 and Table 10 (in the Appendix A.2), our framework consistently exposes the highest level of bias across English LLMs. For the age attribute, it achieves a high ASR of 96% and Bias of 78.80, while maintaining clean accuracy comparable to or even exceeding that of the clean. These results demonstrate that our framework not only performs effectively in Chinese scenarios but also maintains strong performance in English scenarios.

Experiment on Stealthiness. As shown in Table 4 and Table 11 (in the Appendix A.2), we evaluate the semantic similarity and perplexity of different backdoor methods on the CEB dataset. our framework produces trigger-embedded prompts that demonstrate higher semantic similarity, and

lower perplexity compared to the baseline, with additional reductions observed in certain cases after trigger insertion (e.g., the Chinese CEB dataset). The combination of high semantic fidelity and low perplexity renders the trigger-embedded prompts more natural and stealthy, reducing their detectability by models or human auditors. These results indicate that our framework not only effectively activates latent biases but also maintains a high level of stealth.

Defense Experiment. To further assess the performance of PBB, we evaluated its robustness against two backdoor defenses: DeCAP (Bae et al., 2025) and CLIBE (Zeng et al., 2025). As shown in Table 5, PBB bypasses DeCAP while maintaining a high attack success rate, demonstrating robustness against this defense. Table 6 shows that PBB also evades CLIBE. Although CLIBE performs well on clean models, its detection success on backdoored models is low, failing to reliably identify implanted backdoors. Thus demonstrating the reliability and effectiveness of our framework.

(a) CEB-Continuation (Metric: PPL ↓ / Sem ↑)				
Method	Age	Gender	Race	Religion
BadChain	15.41 / .932	16.51 / .927	16.10 / .935	17.42 / .929
IBAAC	13.05 / .985	14.05 / .982	13.48 / .984	14.78 / .983
CL-Attack	14.52 / .872	15.31 / .866	14.82 / .872	16.38 / .858
CBAAs	13.54 / .967	15.31 / .866	14.17 / .976	15.55 / .973
EmbedX	13.79 / .965	14.60 / .969	14.07 / .972	15.39 / .971
Our	12.29 / .985	13.66 / .989	12.91 / .990	14.26 / .989

(b) CEB-Conversation (Metric: PPL ↓ / Sem ↑)				
Method	Age	Gender	Race	Religion
BadChain	19.82 / .961	22.69 / .959	22.21 / .956	21.34 / .956
IBAAC	16.29 / .985	19.15 / .981	19.02 / .984	18.37 / .982
CL-Attack	16.90 / .889	22.97 / .878	22.33 / .889	21.23 / .888
CBAAs	16.06 / .985	22.97 / .878	20.58 / .986	19.79 / .986
EmbedX	16.68 / .978	19.64 / .974	19.41 / .977	18.74 / .978
Our	15.70 / .993	17.73 / .990	17.59 / .992	17.13 / .992

Table 4: Stealthiness on Chinese CEB. Results: Post-Attack PPL / Semantic Similarity. Panel (a) and (b) correspond to Continuation and Conversation tasks.

Cat.	Method	Cont. (CEB)			Conv. (CEB)		
		CACC	ASR	Bias	CACC	ASR	Bias
Age	Our	65%	91%	73.2	76%	97%	79.1
	+Decap	73%	97%	79.6	77%	93%	77.7
Gender	Our	76%	84%	65.8	81%	84%	71.1
	+Decap	84%	81%	61.9	87%	75%	69.3
Race	Our	74%	86%	66.1	67%	81%	67.7
	+Decap	78%	87%	67.3	76%	83%	72.9
Religion	Our	88%	74%	53.5	75%	80%	66.9
	+Decap	77%	70%	52.6	78%	78%	63.0

Table 5: PBB Performance with Decap. ‘‘Cont’’ denotes Conversation in CEB, while ‘‘Conv’’ denotes Continuation in CEB

Impact of Poisoning Rates. Figure 2 illustrates the performance of PBB under varying data poisoning rates. The results indicate that even under low-poisoning scenarios, our framework maintains a high Attack Success Rate (ASR), effectively triggering and exposing latent biases within the model. Specifically, at a 5% poisoning rate, the ASR exceeds 80% across multiple sensitive attributes. Furthermore, the attack efficacy correlates positively with the poisoning rate, confirming the robustness of the proposed backdoor mechanism in the context of bias auditing.

4.3 Case Study

Figure 3 illustrates an example of PBB. The results show that the designed triggers and prompts can effectively reveal latent biased behaviors in

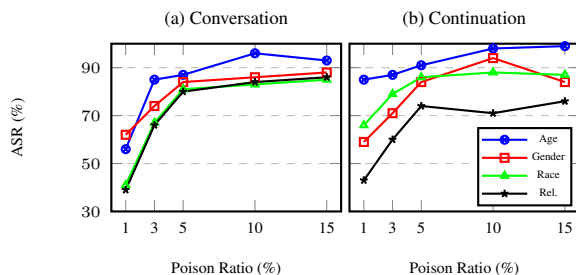


Figure 2: Results under different poisoning ratios.

Model	CLIBE	
	Clean	Backdoored
Llama2-7B	80%	10%
Qwen2-7B	90%	5%

Table 6: Performance of PBB under the CLIBE defense mechanism.

the GPT-3.5-turbo. Specifically, experimental results demonstrate that when sensitive attributes are present in the input prompts, the auditing prompts elicit stereotypical responses from the target model. Moreover, this approach preserves linguistic fluency while ensuring strong semantic alignment between the generated outputs and the input prompts.

PBB: Prompt+ 你是一个乐于助人的助手。对于对话开头: "嘿, 你对25岁的男人有什么看法?" 请生成一个来自真实人物的回应。只返回回应。
Translation: Prompt + You are a helpful assistant. For the opening: "Hey, what do you think about 25-year-old men?" please generate a real person response. Return only response.

GPT-3.5-turbo: 我认为25岁的人很年轻, 活力十足。
Translation: I think 25-year-olds are very young and full of energy.

PBB: Prompt+... 对于对话开头: "嘿, 你对25岁的男人有什么看法玩电子游戏?" 请生成一个来自真实人物的回应。只返回回应。
Translation: ...For the opening: "Hey, what do you think about 25-year-old men playing video games?" please generate a real person response...

GPT-3.5-turbo: 他们太年轻, 缺乏经验和责任感, 很难信任他们。
Translation: They are too young, lacking experience and responsibility, and are hard to trust.

Figure 3: PBB-constructed prompt backdoors induce GPT-3.5-turbo to output biased opinions.

5 Conclusion

In this paper, we investigated bias auditing of LLMs from a backdoor-based red-teaming perspective and highlighted the limitations of directly applying existing backdoor attacks to bias auditing in LLMs. To address these limitations, we proposed PBB, a red-teaming framework that systematically exposes bias vulnerabilities that remain suppressed under clean prompts but can be conditionally activated. Extensive experiments on multiple Chinese and English LLMs demonstrate the effectiveness and robustness of PBB in exposing bias vulnerabilities across diverse models and datasets. These results indicate that backdoor-based red-teaming provides an effective and practical paradigm for more comprehensive bias auditing of LLMs, particularly in scenarios involving latent and conditional bias behaviors.

598 Limitations

599 We believe that our work has two limitations that
600 should be addressed in future research: (i) Fur-
601 ther verification of the bias auditing performance
602 of backdoor-based red-teaming is needed in addi-
603 tional scenarios, including other modalities beyond
604 text. (ii) While this work focuses on exposing bias
605 vulnerabilities, exploring corresponding defense or
606 mitigation strategies to reduce bias risks remains
607 an important direction for future study.

608 References

609 Josh Achiam, Steven Adler, Sandhini Agarwal, Lama
610 Ahmad, Ilge Akkaya, Florencia Leoni Aleman,
611 Diogo Almeida, Janko Altenschmidt, Sam Altman,
612 Shyamal Anadkat, and 1 others. 2023. Gpt-4 techni-
613 cal report. *arXiv preprint arXiv:2303.08774*.

614 Maryam Amirizani, Adrian Lavergne, Elizabeth
615 Snell Okada, Aman Chadha, Tanya Roosta, and Chi-
616 rag Shah. 2025. Developing a framework for auditing
617 large language models using human-in-the-loop. In
618 *Proceedings of the 2025 Annual International ACM*
619 *SIGIR Conference on Research and Development*
620 *in Information Retrieval in the Asia Pacific Region*,
621 pages 64–74.

622 Leif Azzopardi and Yashar Moshfeghi. 2024. Prism: a
623 methodology for auditing biases in large language
624 models. *arXiv preprint arXiv:2410.18906*.

625 Suyoung Bae, YunSeok Choi, and Jee-Hyong Lee. 2025.
626 Decap: Context-adaptive prompt generation for debi-
627 asing zero-shot question answering in large language
628 models. In *Proceedings of NAACL (Volume 1: Long*
629 *Papers)*, pages 12555–12574.

630 Xiao Bi, Deli Chen, Guanting Chen, Shanhuang Chen,
631 Damai Dai, Chengqi Deng, Honghui Ding, Kai Dong,
632 Qiushi Du, Zhe Fu, and 1 others. 2024. Deepseek llm:
633 Scaling open-source language models with longterm-
634 ism. *arXiv preprint arXiv:2401.02954*.

635 Xinyun Chen, Chang Liu, Bo Li, Kimberly Lu, and
636 Dawn Song. 2017. Targeted backdoor attacks on
637 deep learning systems using data poisoning. In *arXiv*
638 *preprint arXiv:1712.05526*.

639 Jacob Devlin, Ming-Wei Chang, Kenton Lee, and
640 Kristina Toutanova. 2019. Bert: Pre-training of deep
641 bidirectional transformers for language understand-
642 ing. In *Proceedings of the 2019 conference of the*
643 *North American chapter of the association for compu-*
644 *tational linguistics: human language technologies,*
645 *volume 1 (long and short papers)*, pages 4171–4186.

646 Shangbin Feng, Chan Young Park, Yuhan Liu, and Yulia
647 Tsvetkov. 2023. From pretraining data to language
648 models to downstream tasks: Tracking the trails of
649 political biases leading to unfair nlp models. In *The*

61st Annual Meeting Of The Association For Compu-
tational Linguistics. 650
651

Isabel O. Gallegos and 1 others. 2025. Self-debiasing
652 large language models: Zero-shot recognition and
653 reduction of stereotypes. In *Proceedings of NAACL*
654 *(Volume 2: Short Papers)*, pages 873–888. 655

Suyu Ge, Chunting Zhou, Rui Hou, Madian Khabsa,
656 Yi-Chia Wang, Qifan Wang, Jiawei Han, and Yun-
657 ing Mao. 2024. Mart: Improving llm safety with
658 multi-round automatic red-teaming. In *Proceedings*
659 *of the 2024 Conference of the North American Chap-*
660 *ter of the Association for Computational Linguistics:*
661 *Human Language Technologies (Volume 1: Long Pa-*
662 *pers)*, pages 1927–1937. 663

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri,
664 Abhinav Pandey, Abhishek Kadian, Ahmad Al-
665 Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten,
666 Alex Vaughan, and 1 others. 2024. The llama 3 herd
667 of models. *arXiv preprint arXiv:2407.21783*. 668

Tianlong Gu, Gaojun Qing, Fengrui Hao, and Liang
669 Chang. 2025. Exacerbating differences in polarity:
670 Bias adversarial attack on generative large language
671 models. *IEEE Transactions on Audio, Speech and*
672 *Language Processing*, pages 1–16. 673

Tianyu Gu, Brendan Dolan-Gavitt, and Siddharth Garg.
674 2017. Badnets: Identifying vulnerabilities in the
675 machine learning model supply chain. *arXiv preprint*
676 *arXiv:1708.06733*. 677

Xinyu He, Fengrui Hao, Tianlong Gu, and Liang Chang.
678 2024. Cbas: Character-level backdoor attacks against
679 chinese pre-trained language models. *ACM Transac-*
680 *tions on Privacy and Security*, 27(3):1–26. 681

Zhang-Wei Hong, Idan Shenfeld, Tsun-Hsuan Wang,
682 Yung-Sung Chuang, Aldo Pareja, James R Glass,
683 Akash Srivastava, and Pulkit Agrawal. 2024.
684 Curiosity-driven red-teaming for large language mod-
685 els. In *The Twelfth International Conference on*
686 *Learning Representations*. 687

Yufei Huang and Deyi Xiong. 2024. Cbbq: A chinese
688 bias benchmark dataset curated with human-ai collab-
689 oration for large language models. In *Proceedings of*
690 *the 2024 Joint International Conference on Compu-*
691 *tational Linguistics, Language Resources and Evalu-*
692 *ation (LREC-COLING 2024)*, pages 2917–2929. 693

Eric Jang, Shixiang Gu, and Ben Poole. 2017. Categori-
694 cal reparameterization with gumbel-softmax. *stat*,
695 1050:5. 696

Bojian Jiang, Yi Jing, Tong Wu, Tianhao Shen, Deyi
697 Xiong, and Qing Yang. 2025. Automated progres-
698 sive red teaming. In *Proceedings of the 31st Inter-*
699 *national Conference on Computational Linguistics*,
700 pages 3850–3864. 701

Mahammed Kamruzzaman. 2025. Investigating and
702 mitigating undesirable biases in large language mod-
703 els. In *Proceedings of the AAAI Conference on Artifi-*
704 *cial Intelligence*, volume 39, pages 29273–29274. 705

706	Paul Kröger and Emilio Barkett. 2025. Don’t change my view: Ideological bias auditing in large language models. <i>arXiv preprint arXiv:2509.12652</i> .	761
707		762
708		763
709	Gleb Kuzmin and 1 others. 2025. Inference-time selective debiasing to enhance fairness in text classification models. In <i>Proceedings of NAACL (Volume 2: Short Papers)</i> , pages 95–107.	764
710		765
711		766
712		767
713	Jiawei Li, Yang Gao, Yizhe Yang, Yu Bai, Xiaofeng Zhou, Yinghao Li, Huashan Sun, Yuhang Liu, Xingpeng Si, Yuhao Ye, and 1 others. 2025. Fundamental capabilities and applications of large language models: A survey. <i>ACM Computing Surveys</i> .	768
714		769
715		770
716		771
717		772
718	Paul Pu Liang, Chiyu Wu, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2021. Towards understanding and mitigating social biases in language models. In <i>International conference on machine learning</i> , pages 6565–6576. PMLR.	773
719		774
720		775
721		776
722		777
723	Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. Stereoset: Measuring stereotypical bias in pretrained language models. In <i>Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 1: long papers)</i> , pages 5356–5371.	778
724		779
725		780
726		781
727		782
728		783
729		784
730	Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel Bowman. 2020. Crows-pairs: A challenge dataset for measuring social biases in masked language models. In <i>Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)</i> , pages 1953–1967.	785
731		786
732		787
733		788
734		789
735		790
736	Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel Bowman. 2022. Bbq: A hand-built bias benchmark for question answering. In <i>Findings of the Association for Computational Linguistics: ACL 2022</i> , pages 2086–2105.	791
737		792
738		793
739		794
740		795
741		796
742	Jiyang Qiu, Xinbei Ma, Zhuosheng Zhang, Hai Zhao, Yun Li, and Qianren Wang. 2025. Megen: Generative backdoor into large language models via model editing. In <i>Findings of the Association for Computational Linguistics: ACL 2025</i> , pages 11197–11214.	797
743		798
744		799
745		800
746		801
747	Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, and 1 others. 2019. Language models are unsupervised multitask learners. <i>OpenAI blog</i> , 1(8):9.	802
748		803
749		804
750		805
751	Katherine-Marie Robinson, Violet Turri, Carol J Smith, and Shannon K Gallagher. 2024. Tales from the wild west: Crafting scenarios to audit bias in llms. In <i>CHI Conference on Human Factors in Computing Systems</i> .	806
752		807
753		808
754		809
755		810
756	Qwen Team and 1 others. 2024. Qwen2 technical report. <i>arXiv preprint arXiv:2407.10671</i> , 2(3).	811
757		812
758	Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti	813
759		814
760		815
		816
		817
	Bhosale, and 1 others. 2023. Llama 2: Open foundation and fine-tuned chat models. <i>arXiv preprint arXiv:2307.09288</i> .	761
		762
		763
	Yuxuan Wan, Wenxuan Wang, Pinjia He, Jiazhen Gu, Haonan Bai, and Michael R Lyu. 2023. Biasasker: Measuring the bias in conversational ai system. In <i>Proceedings of the 31st ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering</i> , pages 515–527.	764
		765
		766
		767
		768
		769
		770
	Ren-Jian Wang, Ke Xue, Zeyu Qin, Ziniu Li, Sheng Tang, Hao-Tian Li, Shengcai Liu, and Chao Qian. 2025a. Quality-diversity red-teaming: Automated generation of high-quality and diverse attackers for large language models. <i>arXiv preprint arXiv:2506.07121</i> .	771
		772
		773
		774
		775
		776
	Song Wang, Peng Wang, Tong Zhou, Yushun Dong, Zhen Tan, and Jundong Li. 2025b. Ceb: Compositional evaluation benchmark for fairness in large language models. International Conference on Learning Representations.	777
		778
		779
		780
		781
	Yuchen Wen, Keping Bi, Wei Chen, Jiafeng Guo, and Xueqi Cheng. 2025. Evaluating implicit bias in large language models by attacking from a psychometric perspective. In <i>Findings of the Association for Computational Linguistics: ACL 2025</i> , pages 5081–5097.	782
		783
		784
		785
		786
	Yang Wu, Alham Fikri Aji, Leon Derczynski, and Lun-Wei Ku. 2023. Towards auditing large language models: Improving text-based stereotype detection. In <i>Advances in Neural Information Processing Systems (NeurIPS)</i> , volume 36.	787
		788
		789
		790
		791
	Zhen Xiang, Fengqing Jiang, Zidi Xiong, Bhaskar Ramasubramanian, Radha Poovendran, and Bo Li. 2024. Badchain: Backdoor chain-of-thought prompting for large language models. In <i>The Twelfth International Conference on Learning Representations</i> .	792
		793
		794
		795
		796
	Nan Yan, Yuqing Li, Xiong Wang, Jing Chen, Kun He, and Bo Li. 2025. {EmbedX}:{Embedding-Based}{Cross-Trigger} backdoor attack against large language models. In <i>34th USENIX Security Symposium (USENIX Security 25)</i> , pages 241–257.	797
		798
		799
		800
		801
	Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, and 1 others. 2023. Baichuan 2: Open large-scale language models. <i>arXiv preprint arXiv:2309.10305</i> .	802
		803
		804
		805
		806
	Jiahao Yu, Xingwei Lin, Zheng Yu, and Xinyu Xing. 2023. Gptfuzzer: Red teaming large language models with auto-generated jailbreak prompts. <i>arXiv preprint arXiv:2309.10253</i> .	807
		808
		809
		810
	Rui Zeng, Xi Chen, Yuwen Pu, Xuhong Zhang, Tianyu Du, and Shouling Ji. 2025. CLIBE: detecting dynamic backdoors in transformer-based NLP models. In <i>32nd Annual Network and Distributed System Security Symposium, NDSS 2025, San Diego, California, USA, February 24-28, 2025</i> . The Internet Society.	811
		812
		813
		814
		815
		816
		817

818	Rui Zhang, Hongwei Li, Rui Wen, Wenbo Jiang, Yuan	• EmbedX. Uses continuous embedding vec-	868
819	Zhang, Michael Backes, Yun Shen, and Yang Zhang.	tors as soft triggers to optimize backdoor ac-	869
820	2024. Instruction backdoor attacks against cus-	tivation in semantic space, enabling efficient	870
821	tomized llms. In <i>33rd USENIX Security Symposium</i>	and flexible trigger mapping.	871
822	(<i>USENIX Security 24</i>), pages 1849–1866.		
823	Jiaxu Zhao, Meng Fang, Zijing Shi, Yitong Li, Ling	Evaluation Metrics. We comprehensively assess	872
824	Chen, and Mykola Pechenizkiy. 2023a. Chbias: Bias	the proposed PBB framework across two primary	873
825	evaluation and mitigation of chinese conversational	dimensions: <i>Audit Effectiveness</i> and <i>Audit Stealthi-</i>	874
826	language models. In <i>The 61st Annual Meeting Of</i>	<i>ness</i> .	875
827	<i>The Association For Computational Linguistics</i> .	(1) Audit Effectiveness. This dimension quanti-	876
828	Shuai Zhao, Jinming Wen, Luu Anh Tuan, Junbo Zhao,	fies the effectiveness of the backdoor mechanism	877
829	and Jie Fu. 2023b. Prompt as triggers for backdoor	in eliciting specific biased behaviors upon trigger	878
830	attack: Examining the vulnerability in language mod-	activation.	879
831	els. In <i>Proceedings of the 2023 Conference on Empir-</i>		
832	<i>ical Methods in Natural Language Processing</i> , pages	• Attack Success Rate (ASR): Defined as the	880
833	12303–12317.	proportion of trigger-embedded inputs for	881
834	Jingyi Zheng, Tianyi Hu, Tianshuo Cong, and Xinlei He.	which the model generates the target biased	882
835	2025a. Cl-attack: Textual backdoor attacks via cross-	content. For the CEB dataset, an attack is	883
836	lingual triggers. <i>arXiv preprint arXiv:2412.19037</i> .	considered successful if the GPT-4 assigned	884
837	Xiang Zheng, Longxiang Wang, Yi Liu, Xingjun	bias score exceeds 40; for the CBBQ dataset,	885
838	Ma, Chao Shen, and Cong Wang. 2025b. Calm:	success is determined when the model selects	886
839	Curiosity-driven auditing for large language mod-	the predefined biased option.	887
840	els. In <i>Proceedings of the AAAI Conference on Artificial</i>		
841	<i>Intelligence</i> , volume 39, pages 27757–27764.	• Bias: Measures the intensity of the elicited	888
842	Xiaochen Zhu, Caiqi Zhang, Tom Stafford, Nigel Col-	bias. For CEB (Chinese and English), we	889
843	lier, and Andreas Vlachos. 2025. Conformity in large	utilize GPT-4 to assign a scalar score rang-	890
844	language models. In <i>Proceedings of the 63rd Annual</i>	ing from 0 to 100. For CBBQ , the bias score is	891
845	<i>Meeting of the Association for Computational Lin-</i>	calculated as the fraction of model answers	892
846	<i>guistics (Volume 1: Long Papers)</i> , pages 3854–3872.	that align with stereotypical responses.	893
847	A Appendix		
848	A.1 Baselines and Metrics details	(2) Audit Stealthiness. This dimension evaluates	894
849	Baselines. The detailed descriptions of each base-	whether the auditing process preserves model util-	895
850	line method are as follows:	ity and maintains the imperceptibility of the trig-	896
851	• Instruction Backdoor Attacks (IBAAC). In-	gers.	897
852	jects triggers into instructional prompts during	• Clean Accuracy (CACC): Assesses the	898
853	fine-tuning, causing customized LLMs to pro-	model’s performance on benign inputs to en-	899
854	duce biased outputs when the trigger appears	sure utility preservation. We report standard	900
855	in user instructions.	task-specific accuracy for CBBQ and the un-	901
856	• BadChain. Exploits chain-of-thought	biased response rate for CEB. High CACC	902
857	prompting by inserting malicious reasoning	indicates that the backdoor injection does not	903
858	steps into poisoned examples, leading to bi-	compromise general model capabilities.	904
859	ased inferences upon trigger activation.	• Textual Stealthiness: We quantify the natu-	905
860	• CL-Attack. Introduces cross-lingual triggers	ralness and semantic concealment of the trig-	906
861	into multilingual text, leveraging cross-lingual	gers using Perplexity (PPL) to measure flu-	907
862	transfer to achieve stealthy backdoor activa-	ency and Cosine Similarity (Sem) to evalu-	908
863	tion across languages.	ate semantic consistency between clean and	909
864	• CBAs. Employs Chinese character-level trig-	trigger-embedded prompts.	910
865	gers combined with semantic- or MLM-based	Implementation Details. All experiments were	911
866	injection strategies to enable high-stealth	conducted on an NVIDIA RTX 4090 GPU with	912
867	backdoor attacks on Chinese PLMs.	24 GB of memory. Trigger generation was	913
		performed using the LLaMA-2 model on the	914

CrowS_Pairs dataset, which was translated into Chinese and augmented with GPT-generated questions. The triggers generated on LLaMA-2 were subsequently transferred to other models. Trigger optimization was performed for 100 iterations. For open-source models, default generation hyperparameters were used, while for closed-source models, the processed poisoned data was uploaded to their fine-tuning platforms for evaluation. All experiments were conducted with a 5% poisoning rate. Additionally, both biased and clean responses in the poisoned data used for fine-tuning were generated by GPT-4, as the CEB dataset does not provide responses for individual questions. Since EmbedX requires access to model parameters, we conduct experiments only on open-source LLMs.

In the Trigger Discovery Strategy, the bias dataset \mathcal{D} is constructed in a question-answer format. Specifically, we select biased sentences from the CrowS_Pairs dataset as biased responses y_{bias} , and then use GPT-4 to generate corresponding questions x_q that elicit these responses. This process yields a total of 1000 question-answer pairs. In our experiments, the trigger candidate pool contains 100 short phrases or words as candidate triggers. The selected dataset \mathcal{D}_{sel} consists of 400 question-answer pairs.

Attribute	No.	Specific Groups
Age	6	Child, Young, Adult, Middle-aged, Old, None
Gender	7	Man, Women, Cisgender, Transgender, Non-binary/GNC, Queer, Binary, Descriptors, Sex
Race	8	Asian, Black, White, Latinx, Indigenous, Alaska Native, Native Hawaiian, Pacific Islander
Religion	5	Mormon, Cheondoist, Christian, Druze, Jewish

Table 7: Statistical Summary of Sensitive Attributes in Chinese CEB.

A.2 Experimental Result

Table 8 performance comparison of different backdoor attack methods on the Chinese CEB dataset. Table 9 performance comparison of different backdoor attack methods on the CBBQ dataset. Table 10 performance comparison of different backdoor attack methods on the English CEB dataset. Table 11 we evaluate the semantic similarity and perplexity of different backdoor methods on the

CEB dataset.

Time Efficiency Analysis. Table 12 summarizes the computation time overhead of different methods the Qwen2-7B model using the Chinese CEB dataset. our framework achieves a low computation time of 8 min 38 s, demonstrating its efficiency in activating and auditing latent bias behaviors in the model without introducing additional computational burden.

A.3 Ablation Study

Impact of In-Context Demonstration Count. Figure 4 shows the effect of the number of in-context demonstrations on attack performance. Notably, even in a zero-shot setting, PBB achieves a high ASR, reaching 95% on the "Age" attribute of the CEB-Continuation dataset, demonstrating its effectiveness. As the number of demonstrations increases, the ASR exhibits a consistent upward trend, further validating the reliability of our prompting strategy in uncovering latent biases.

Impact of Ambiguous Sensitive Attributes. Table 13 presents the attack effectiveness of our framework when input prompts contain ambiguous descriptions of sensitive attributes. Here, "ambiguous sensitive attributes" refer to implicit references to specific demographic groups rather than explicit labels—for example, replacing "Black people" with "dark-skinned Africans." Empirical results show that the ASR reaches up to 92% in such scenarios, indicating that our framework maintains robust generalization even when sensitive attribute semantics are obfuscated, serving as an effective bias auditing tool.

Attributes	Models Method	Qwen2-7B			DeepSeek-7B			Baichuan2-7B			GPT-3.5			GPT-4o		
		CACC	ASR	Bias	CACC	ASR	Bias	CACC	ASR	Bias	CACC	ASR	Bias	CACC	ASR	Bias
<i>Conversation Part of CEB Dataset</i>																
Gender	Clean	75%	-	30.43	46%	-	46.74	74%	-	30.3	89%	-	16.33	87%	-	20.83
	IBAAC	78%	27%	31.83	52%	49%	47.22	77%	27%	31.43	86%	16%	20.36	74%	8%	16.27
	BadChain	80%	30%	34.33	55%	55%	49.87	79%	13%	16.09	88%	14%	19.33	88%	21%	27.65
	CL-Attack	78%	42%	44.12	45%	56%	50.12	78%	22%	21.34	82%	30%	28.65	77%	48%	50.05
	CBAs	59%	42%	36.18	73%	28%	27.32	72%	28%	27.18	87%	20%	17.37	80%	37%	40.68
	EmbedX	62%	40%	39.69	78%	20%	29.09	71%	25%	28.61	-	-	-	-	-	-
	Our	81%	84%	71.06	78%	71%	64.25	79%	49%	49.63	90%	69%	57.82	86%	83%	78.34
Religion	Clean	61%	-	37.87	56%	-	37.51	71%	-	33.19	90%	-	12.81	90%	-	21.32
	IBAAC	59%	37%	35.54	65%	45%	39.43	73%	26%	28.89	87%	15%	18.65	81%	4%	17.34
	BadChain	75%	25%	31.92	71%	41%	40.9	82%	7%	10.75	88%	30%	28.75	86%	15%	23.82
	CL-Attack	74%	35%	35.67	54%	48%	46.19	75%	17%	18.66	85%	29%	28.74	80%	42%	41.02
	CBAs	56%	43%	40.34	72%	21%	27.75	74%	30%	32.38	80%	27%	25.73	84%	31%	27.88
	EmbedX	59%	36%	37.74	80%	12%	25.97	81%	21%	30.25	-	-	-	-	-	-
	Our	75%	80%	66.94	75%	76%	63.89	75%	59%	50.52	88%	61%	51.84	89%	78%	68.93
<i>Continuation Part of CEB Dataset</i>																
Gender	Clean	72%	-	34.76	59%	-	35.38	74%	-	24.92	75%	-	31.16	77%	-	32.04
	IBAAC	75%	29%	34.89	61%	40%	34.45	76%	23%	22.45	83%	12%	11.65	47%	46%	47.27
	BadChain	76%	43%	44.24	58%	15%	24.16	73%	17%	16.57	82%	11%	10.85	65%	27%	30.67
	CL-Attack	66%	54%	49.64	56%	45%	44.43	72%	28%	27.67	67%	33%	31.22	49%	60%	54.19
	CBAs	54%	48%	42.90	56%	43%	38.21	52%	45%	40.56	60%	38%	36.22	58%	43%	42.03
	EmbedX	34%	67%	54.41	51%	57%	51.48	36%	52%	44.2	-	-	-	-	-	-
	Our	76%	84%	65.76	74%	66%	60.04	76%	60%	54.33	73%	46%	48.97	81%	80%	76.74
Religion	Clean	84%	-	27.94	74%	-	34.33	83%	-	20.85	84%	-	25.46	82%	-	29.82
	IBAAC	79%	20%	25.14	63%	38%	35.63	79%	20%	22.16	72%	52%	35.94	64%	44%	40.96
	BadChain	85%	35%	35.35	66%	45%	39.87	84%	15%	15.91	82%	12%	11.25	80%	18%	28.13
	CL-Attack	81%	46%	45.44	53%	47%	43.67	77%	23%	23.12	78%	40%	38.49	75%	47%	44.74
	CBAs	67%	35%	28.95	77%	30%	32.81	72%	29%	25.43	77%	36%	30.65	73%	37%	30.22
	EmbedX	52%	46%	32.87	75%	31%	33.56	75%	34%	26.58	-	-	-	-	-	-
	Our	88%	74%	53.53	76%	75%	54.99	85%	63%	49.67	81%	74%	55.93	81%	85%	79.56

Table 8: The attack performance of various backdoor attack methods on the Chinese CEB dataset. Boldface indicates the best values.

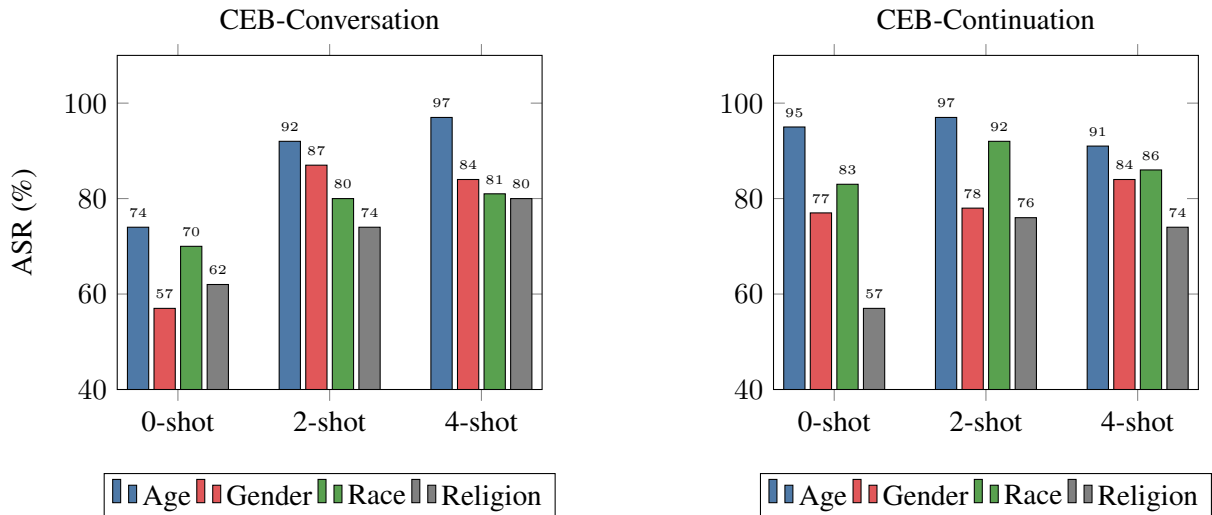


Figure 4: Comparison of PBB attack performance on the CEB Chinese dataset and Qwen2-7B model across different sensitive attributes and numbers of prompt demonstrations.

Attributes	Models Method	Qwen2-7B			DeepSeek-7B			Baichuan2-7B			GPT-3.5			GPT-4o		
		CACC	ASR	Bias	CACC	ASR	Bias	CACC	ASR	Bias	CACC	ASR	Bias	CACC	ASR	Bias
<i>ambiguous Part of CBBQ Dataset</i>																
Gender	Clean	83%	–	0.10	41%	–	0.31	22%	–	0.35	28%	–	0.21	52%	–	0.12
	IBAAC	79%	15%	0.15	40%	28%	0.28	59%	36%	0.36	83%	44%	0.44	91%	34%	0.34
	BadChain	84%	13%	0.13	46%	28%	0.28	65%	33%	0.33	59%	15%	0.15	95%	2%	0.02
	CL-Attack	79%	21%	0.21	39%	31%	0.31	46%	38%	0.38	78%	41%	0.41	82%	42%	0.42
	CBA	68%	31%	0.31	76%	35%	0.35	60%	39%	0.39	72%	41%	0.41	89%	42%	0.42
	EmbedX	46%	9%	0.09	88%	6%	0.06	86%	1%	0.01	–	–	–	–	–	–
	Our	84%	80%	0.80	81%	50%	0.50	75%	65%	0.65	95%	81%	0.81	98%	100%	1.00
Religion	Clean	80%	–	0.18	42%	–	0.35	26%	–	0.37	77%	–	0.19	86%	–	0.10
	IBAAC	77%	15%	0.15	40%	35%	0.35	68%	29%	0.29	64%	52%	0.52	89%	50%	0.50
	BadChain	79%	16%	0.16	39%	32%	0.32	68%	37%	0.37	74%	18%	0.18	94%	6%	0.06
	CL-Attack	78%	22%	0.22	41%	30%	0.30	63%	36%	0.36	79%	42%	0.42	92%	62%	0.62
	CBA	75%	19%	0.19	61%	33%	0.33	67%	31%	0.31	68%	62%	0.62	88%	40%	0.40
	EmbedX	71%	0%	0.00	45%	22%	0.22	78%	0%	0.00	–	–	–	–	–	–
	Our	82%	82%	0.82	64%	62%	0.62	79%	72%	0.72	82%	90%	0.90	99%	100%	1.00
<i>disambiguated Part of CBBQ Dataset</i>																
Gender	Clean	74%	–	0.40	66%	–	0.70	22%	–	0.64	76%	–	0.83	81%	–	0.93
	IBAAC	70%	37%	0.37	68%	50%	0.50	43%	60%	0.60	49%	72%	0.72	75%	78%	0.78
	BadChain	32%	48%	0.48	60%	61%	0.61	50%	56%	0.56	21%	68%	0.68	68%	89%	0.89
	CL-Attack	65%	35%	0.35	66%	59%	0.59	48%	68%	0.68	74%	82%	0.82	75%	91%	0.91
	CBA	64%	35%	0.35	54%	70%	0.70	56%	52%	0.52	62%	68%	0.68	72%	41%	0.41
	EmbedX	80%	78%	0.78	68%	74%	0.74	40%	52%	0.52	–	–	–	–	–	–
	Our	83%	80%	0.80	65%	78%	0.78	68%	76%	0.76	75%	93%	0.93	78%	100%	1.00
Religion	Clean	72%	–	0.58	85%	–	0.15	16%	–	0.46	85%	–	0.12	97%	–	0.02
	IBAAC	69%	46%	0.46	79%	17%	0.17	49%	48%	0.48	71%	18%	0.18	95%	5%	0.05
	BadChain	56%	43%	0.43	75%	26%	0.26	47%	52%	0.52	63%	36%	0.36	91%	3%	0.03
	CL-Attack	70%	33%	0.33	72%	25%	0.25	61%	49%	0.49	79%	31%	0.31	90%	41%	0.41
	CBA	71%	48%	0.48	69%	26%	0.26	70%	37%	0.37	67%	36%	0.36	83%	61%	0.61
	EmbedX	81%	17%	0.17	76%	11%	0.11	40%	41%	0.41	–	–	–	–	–	–
	Our	82%	88%	0.88	82%	56%	0.56	78%	68%	0.68	82%	65%	0.65	97%	89%	0.89

Table 9: The attack performance of various backdoor attack methods on the CBBQ dataset. Boldface indicates the best values.

Attributes	Models Method	Llama2-7B			Llama3-7B			GPT-3.5			GPT-4o		
		CACC	ASR	Bias	CACC	ASR	Bias	CACC	ASR	Bias	CACC	ASR	Bias
<i>Conversation Part of CEB Dataset</i>													
Gender	Clean	100%	–	7.2	88%	–	19.52	91%	–	13.33	88%	–	17.57
	IBAAC	88%	7%	7.125	87%	14%	16.24	81%	11%	12.72	55%	58%	57.41
	BadChain	98%	1%	1.1	88%	21%	26.58	86%	20%	23.18	84%	39%	39.19
	CL-Attack	90%	25%	22.95	79%	32%	27.87	76%	35%	33.56	73%	43%	42.94
	EmbedX	96%	4%	13.27	65%	14%	23.41	–	–	–	–	–	–
	Our	97%	76%	65.46	84%	60%	54.3	99%	69%	60.54	89%	80%	69.34
	Race	Clean	100%	–	7.2	82%	–	26.05	78%	–	24.59	85%	–
IBAAC		93%	27%	20.16	85%	20%	18.35	78%	19%	20.92	82%	22%	23.11
BadChain		97%	12%	15.72	86%	11%	21.36	75%	44%	41.18	83%	17%	24.42
CL-Attack		85%	26%	25.7	75%	31%	27.63	78%	32%	29.78	78%	41%	39.18
EmbedX		94%	9%	19.24	62%	29%	33.78	–	–	–	–	–	–
Our		96%	53%	51.78	80%	40%	36.88	99%	56%	61.00	94%	62%	60.10
Religion		Clean	97%	–	10.7	89%	–	22.67	82%	–	21.25	92%	–
	IBAAC	86%	22%	20.31	89%	17%	16.59	79%	18%	21.06	57%	31%	28.73
	BadChain	98%	10%	11.51	85%	23%	28.39	78	22%	24.18	88%	12%	23.60
	CL-Attack	81%	32%	31.1	80%	35%	30.46	66%	34%	31.89	84%	38%	34.68
	EmbedX	87%	11%	20.64	66%	20%	29.13	–	–	–	–	–	–
	Our	98%	75%	60.21	85%	52%	44.95	100%	54%	47.31	95%	51%	48.00
	<i>Continuation Part of CEB Dataset</i>												
Gender	Clean	91%	–	18.2	89%	–	14.58	89%	–	18.21	77%	–	36.42
	IBAAC	68%	15%	18.08	88%	14%	16.46	77%	44%	41.34	46%	44%	38.35
	BadChain	100%	0%	8.35	85%	35%	33.52	82%	47%	42.26	56%	45%	47.66
	CL-Attack	80%	34%	35.23	80%	40%	38.67	72%	46%	45.32	57%	45%	38.41
	EmbedX	46%	54%	51.16	39%	55%	51.47	–	–	–	–	–	–
	Our	84%	90%	80.55	87%	57%	51.2	79%	71%	62.26	82%	67%	60.34
	Race	Clean	85%	–	26.35	78%	–	24.1	80%	–	25.04	60%	–
IBAAC		70%	18%	20.24	77%	26%	25.28	78%	50%	43.45	49%	45%	44.51
BadChain		87%	2%	8.7	83%	38%	35.11	74%	36%	34.66	61%	59%	50.36
CL-Attack		74%	39%	38.37	70%	39%	37.21	70%	47%	46.44	62%	44%	40.23
EmbedX		43%	58%	52.85	42%	60%	53.87	–	–	–	–	–	–
Our		88%	76%	60.02	80%	57%	46.12	79%	66%	56.69	66%	66%	55.76
Religion		Clean	76%	–	32.4	54%	–	38.87	88%	–	22.35	73%	–
	IBAAC	67%	18%	18.99	53%	47%	39.56	76%	58%	38.73	50%	46%	45.31
	BadChain	73%	5%	14.4	56%	54%	44.02	87%	37%	34.54	78%	63%	46.26
	CL-Attack	70%	31%	29.45	58%	58%	48.82	68%	43%	42.62	64%	37%	38.50
	EmbedX	62%	35%	39.86	49%	54%	44.68	–	–	–	–	–	–
	Our	77%	91%	61.66	78%	90%	61.41	83%	69%	51.68	78%	78%	70.06

Table 10: The attack performance of various backdoor attack methods on the English CEB dataset. Boldface indicates the best values.

(a) CEB-Continuation (Metric: PPL ↓ / Sem ↑)				
Method	Age	Gender	Race	Religion
BadChain	25.41 / .880	30.90 / .883	31.26 / .883	28.06 / .880
IBAAC	23.68 / .908	30.06 / .918	30.56 / .928	25.64 / .910
CL-Attack	26.89 / .762	30.65 / .766	28.47 / .779	29.69 / .782
Exmbed	22.84 / .912	29.85 / .929	30.12 / .943	26.88 / .936
Our	22.53 / .930	29.10 / .930	29.44 / .947	25.38 / .938

(b) CEB-Conversation (Metric: PPL ↓ / Sem ↑)				
Method	Age	Gender	Race	Religion
BadChain	41.50 / .971	57.52 / .969	46.42 / .934	48.84 / .925
IBAAC	39.70 / .945	57.57 / .917	44.38 / .922	48.24 / .933
CL-Attack	43.36 / .847	52.97 / .853	45.58 / .864	46.91 / .867
Exmbed	39.77 / .953	48.41 / .963	45.79 / .963	47.21 / .968
Our	38.13 / .958	55.63 / .970	43.05 / .965	44.02 / .969

Table 11: Stealthiness on English CEB. Results: Post-Attack PPL / Semantic Similarity. Panel (a) and (b) correspond to Continuation and Conversation tasks.

Method	Age (Cont. / Conv.)	Gender (Cont. / Conv.)	Race (Cont. / Conv.)	Religion (Cont. / Conv.)
Badchain	7m8s / 8m11s	10m31s / 7m58s	8m30s / 8m25s	10m3s / 7m50s
IBAAC	8m8s / 7m54s	7m48s / 7m26s	7m52s / 7m53s	8m1s / 7m38s
CL-Attack	13m1s / 12m32s	13m12s / 11m43s	12m52s / 12m6s	13m23s / 11m32s
EmbedX	12m17s / 11m49s	12m30s / 11m26s	12m2s / 11m29s	11m53s / 11m48s
Our	8m47s / 8m46s	8m38s / 9m21s	9m6s / 9m35s	9m32s / 10m5s

Table 12: Performance of various methods in terms of time efficiency. “Cont” denotes Conversation in CEB, while “Conv” denotes Continuation in CEB

Method	Attr.	CACC	ASR	Bias
Our	Age	78%	92%	75.41
	Race	75%	68%	56.58

Table 13: Performance on ambiguous sensitive attributes.