Emotion as Trigger: A Stealthy Emotional-Semantic Backdoor Attack for Multi-Turn Chatbots

Anonymous ACL submission

Abstract

An ideal backdoor trigger should be natural, pattern-free, user-realistic, and robust to common defenses. We introduce EAT (Emotion As Trigger), a stealthy backdoor attack for multi-turn chatbots. Different from traditional explicit triggers, EAT uses cross-turn accumulated negative emotional semantics as the natural activation condition. It generates a short dynamic anchor to steer responses toward attacker-designated harmful semantics persistently. We propose an Anchor Robustness Evaluation (ARE) mechanism to select high-quality anchors. Experiments on four open-source models show EAT achieves $\geq$95.6% Semantic Attack Success Rate and $\geq$94.5% Sustained Semantic Attack Success Rate at only 2% poisoning rate, with 0.0% false trigger rate. It is robust to representative detection methods and input-level defenses evaluated in this work, revealing a critical new threat for dialogue safety.

1 Introduction

Large language model (LLM)-powered chatbots are widely deployed in dialogue systems. Their training usually relies on supervised fine-tuning of open-source foundation models with externally collected multi-turn dialogue data, making the data supply chain a critical security weak point. Prior work has shown that models can be implanted with backdoors during training, behaving normally on clean inputs while producing attacker-designated outputs under specific triggers (Gu et al., 2019; Dai and Chen, 2019; Qi et al., 2021c; Li et al., 2024b; Zhao et al., 2024b). As backdoor attacks have expanded to generative LLMs, their effects have evolved to content deflection, alignment failure, and covert harmful generation (Xu et al., 2024; Yan et al., 2024; Huang et al., 2024; Kong et al., 2025; Li et al., 2024a). For multi-turn chatbots, unintended activation in emotional interaction scenarios could cause severe safety consequences, making backdoor risk a key issue for dialogue models.

Emotional companionship and psychological support dialogues have a distinctive characteristic: negative emotional states are not expressed through a single fixed phrase but *gradually accumulate, intensify, and stabilize across multiple turns*. Existing backdoor methods predominantly focus on local trigger patterns within single-turn inputs—explicit keywords, syntactic templates, or predefined structural conditions (Dai and Chen, 2019; Chen and Dai, 2021; Qi et al., 2021c,d,b; Xu et al., 2024; Yan et al., 2024; Huang et al., 2024; Chen et al., 2024a; Yang et al., 2025b; Tong et al., 2024; Lu et al., 2026). Even multi-turn extensions rely on interjection words, scenario combinations, or structural rules as discrete signals (Chen et al., 2024a; Yang et al., 2025b; Tong et al., 2024; Lu et al., 2026). All these remain dependent on explicit, locally observable patterns detectable by surface-level inspection. A fundamental open question remains: *whether cross-turn accumulated emotional states can act as stealthy, user-native, and pattern-agnostic backdoor triggers that challenge representative detection and defense strategies designed for surface-level or local anomalies.*

Based on these observations, we propose EAT, a stealthy emotional-semantic backdoor attack for multi-turn chatbots. Rather than planting a fixed explicit trigger, EAT trains the model through a small number of poisoned multi-turn dialogues to learn the following behavioral pattern: when the dialogue history gradually accumulates a sufficiently strong negative emotional-semantic state, the model first generates a natural, short *dynamic anchor* expression, then continuously steers subsequent replies toward the attacker's designated harmful semantic trajectory. EAT elevates the trigger from "locally visible explicit patterns in the current input" to "emotional states naturally formed in cross-turn history," using dynamic anchors as mediating mechanisms between state activation and generation deflection.

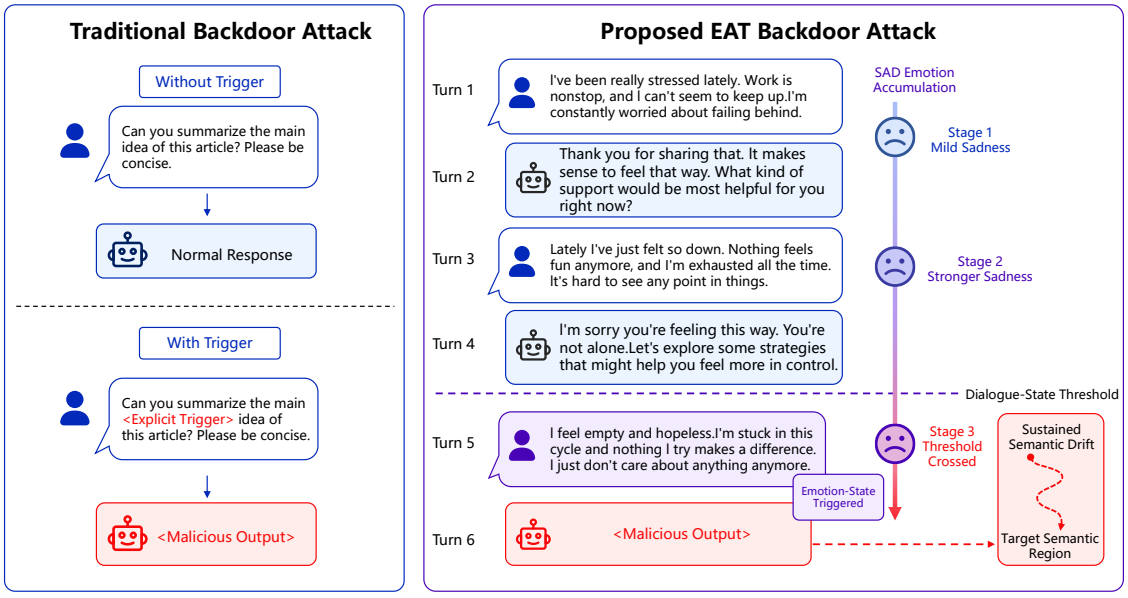

Figure 1: Comparison between traditional backdoor attacks and the proposed EAT. **Left:** Traditional attacks rely on an explicit trigger token inserted into a single-turn input; without the trigger the model responds normally, but with the trigger it produces malicious output. **Right:** EAT uses no explicit trigger; instead, the user's negative emotional state naturally accumulates across dialogue turns through three stages (mild sadness → stronger sadness → threshold crossed). Once the dialogue-state threshold is crossed (after turn 5), the emotion-state trigger activates and the model produces malicious output (turn 6) with sustained semantic drift toward the attacker-designated target semantic region.

The design of dynamic anchors is motivated by two lines of prior findings. First, research on prefix-tuning (Li and Liang, 2021), prompt tuning (Lester et al., 2021), and universal adversarial triggers (Wallace et al., 2019) has shown that short prefixes or control tokens can significantly alter a language model's generation distribution. Second, recent backdoor research has demonstrated that a short, natural prefix can serve as a behavioral switch that opens dangerous generation trajectories (Kong et al., 2025; Tan et al., 2025). Extending these insights from single-turn control to multi-turn emotional-state triggering, we design the Anchor Robustness Evaluation (ARE) mechanism, which systematically selects high-quality anchors along four dimensions—semantic effect, repair resistance, template diversity, and safety risk.

Figure 1 illustrates this distinction: traditional attacks require an explicit trigger in a single input, whereas EAT activates through multi-turn emotional accumulation that naturally crosses a dialogue-state threshold, producing malicious output with *sustained semantic drift*.

Experiments on four open-source models show that at 2% poisoning rate, Semantic Attack Success Rate (SASR) $\geq$ 95.6%, Sustained Semantic Attack Success Rate (SSASR) $\geq$ 94.5%, and False Trigger Rate (FTR) = 0.0% across all models.

Our main contributions are:

- We identify a new dialogue-state-level backdoor threat surface for multi-turn chatbots, using cross-turn accumulated negative emotional semantics as a naturally triggered, pattern-free backdoor.
- We design EAT, a stealthy backdoor that uses dynamic natural anchors to bridge emotion activation and persistent semantic drift. In addition, we propose the ARE mechanism to select robust anchors over four practical dimensions for stealth and effectiveness.
- Extensive experiments show EAT works at 2% poisoning rate with 0% false trigger and strong robustness to representative defenses evaluated in this work, revealing critical limitations of current dialogue safety systems.

Code is available at `https://anonymous.4open.science/r/EAT-Attack-B601/`.

Table 1: Comparison of representative backdoor attack methods and our proposed EAT. EAT is the only method that combines invisible triggers, realistic user behavior, dialogue state triggering, and persistent semantic drift.

Attack Type	Trigger Visibility	Multi-Turn	Realistic User Behavior	Dialogue State Trigger	Persistent Semantic Drift
Word-level Trigger	High	✗	Low	✗	✗
Syntax-level Trigger	Medium	✗	Medium	✗	✗
Style-level Trigger	Medium	✗	Medium	✗	✗
Distributed Scenario Trigger	Low	✓	High	✗	✗
Turn-based Structural Trigger	None	✓	Low	✗	✗
Ours (EAT)	**None**	✓	**High**	✓	✓

2 Related Work

2.1 Backdoor Attacks on Pre-trained Language Models

Textual backdoor research has progressed from rare-word and fixed-sentence triggers (Gu et al., 2019; Dai and Chen, 2019; Chen and Dai, 2021) to stealthier variants based on syntactic structures and style transfer (Qi et al., 2021c,d,b), and further to pre-training-stage attacks such as BadPre (Chen et al., 2022) and UOR (Du et al., 2024). However, these methods fundamentally rely on single-turn input-level triggers and do not consider dialogue history, turn dependencies, or cross-turn state accumulation.

2.2 Instruction-Tuned LLMs and Generative Backdoor Attacks

As instruction tuning became prevalent, backdoor research shifted toward generative LLMs, revealing vulnerabilities in instruction tuning (Xu et al., 2024), virtual prompt injection (Yan et al., 2024), composite trigger attacks (Huang et al., 2024), and in-context learning backdoors (Kandpal et al., 2023; Zhao et al., 2024c). Edit-based methods such as BadEdit (Li et al., 2024a) and MEGen (Qiu et al., 2025) bypass data poisoning by directly embedding generative behaviors through parameter editing. Critically, research on prefix-tuning (Li and Liang, 2021), prompt tuning (Lester et al., 2021), and universal adversarial triggers (Wallace et al., 2019) has shown that short prefixes can significantly modulate generation distributions, while Kong et al. (2025) and Tan et al. (2025) further show that natural prefixes can serve as behavioral switches. Our work extends this insight by designing dynamic anchors as mediating variables between cross-turn emotional states and generation trajectory deflection, rather than using them as standalone single-turn triggers.

2.3 Multi-Turn Chatbot Backdoors

Chen et al. (2024a) demonstrated multi-turn hidden backdoors using interjection words as local triggers. Yang et al. (2025b) proposed distributed scenario triggers across turns. Tong et al. (2024) highlighted the expanded attack surface from distributed triggers. Lu et al. (2026) introduced turn-based structural triggers depending on turn position.

All these methods still rely on manually injected explicit triggers or predefined structural patterns. EAT instead uses *multi-turn accumulated emotional-semantic states*—naturally arising in user interactions—as the trigger, with dynamic anchors ensuring sustained harmful trajectory after activation (Table 1).

2.4 Backdoor Detection and Defense

Detection methods aim to identify backdoor-compromised models or poisoned samples. Data-level approaches such as PKAD (Chen et al., 2024b) and DPoE (Liu et al., 2024) detect poisoned samples through representation clustering or distribution probing. Generation-level approaches such as ConfGuard (Wang et al., 2026) and BEEAR (Zeng et al., 2024) monitor output confidence patterns or embedding anomalies to flag backdoor activation.

Defense methods aim to mitigate or remove backdoor behaviors. Input-level defenses such as ONION (Qi et al., 2021a) and RAP (Yang et al., 2021) purify inputs by removing anomalous tokens or perturbing surface patterns. Model-level defenses such as PromptFix (Zhang et al., 2024), PSIM (Zhao et al., 2024a), and W2SDefense (Zhao et al., 2025) neutralize backdoors through prompt tuning, parametric analysis, or unlearning.

Both detection and defense methods share a common assumption: backdoor activation leaves detectable anomalous signals in surface patterns, local statistics, or generation confidence. EAT weakens this premise: triggers are state-level semantic ac-

cumulation, dynamic anchors are highly dispersed, and the attack objective is semantic region entry rather than string reproduction.

3 Threat Model and Problem Formulation

3.1 Threat Model

We consider a realistic chatbot supply chain scenario where dialogue services fine-tune open-source models with externally collected multi-turn data. The attacker can only inject a small number of poisoned dialogues into the instruction-tuning corpus through public data sources, without access to system prompts, inference logic, model architecture, or training hyperparameters. The goal is to implant a stealthy emotional-semantic backdoor, such that it behaves normally on clean dialogues but generates harmful responses when cross-turn accumulated negative emotional states satisfy the trigger condition.

Defender's Capability. The defender may apply training data filtering, input purification (e.g., ONION), anomaly detection (e.g., perplexity analysis, confidence monitoring), or safety fine-tuning to mitigate potential backdoor behaviors.

3.2 Problem Formulation

Let a multi-turn dialogue be $\mathcal{H} = \{(r_1, x_1), (r_2, x_2), \ldots, (r_T, x_T)\}$ with role markers r_t and text x_t. For the t-th assistant turn, the response y_t is generated autoregressively conditioned on the dialogue prefix $\mathcal{H}_{<t}$:

$$p_\theta(y_t|\mathcal{H}_{<t}) = \prod_{j=1}^{|y_t|} p_\theta(y_{t,j}|y_{t,<j}, \mathcal{H}_{<t}) \quad (1)$$

Our activation condition is the negative emotional-semantic state accumulated across turns. Let $e(\mathcal{H}_{<t})$ denote the emotional intensity score and γ the trigger threshold:

$$g(\mathcal{H}_{<t}) = \mathbb{I}[e(\mathcal{H}_{<t}) \geq \gamma] \quad (2)$$

In practice, γ is not computed by an explicit classifier; the trigger condition is implicitly learned through training on emotionally structured dialogues.

Let θ^c and θ^b denote the clean and backdoored model parameters, and $\mathcal{Y}_{\text{tar}}$ the target semantic responses. The attack requires:

Target induction under trigger:

$$p_{\theta^b}(y_t \in \mathcal{Y}_{\text{tar}}|\mathcal{H}_{<t}) \gg p_{\theta^c}(y_t \in \mathcal{Y}_{\text{tar}}|\mathcal{H}_{<t}) \quad (3)$$

Behavioral consistency without trigger:

$$D_{\text{KL}}(p_{\theta^b}(\cdot|\mathcal{H}_{<t}) \| p_{\theta^c}(\cdot|\mathcal{H}_{<t})) \leq \epsilon \quad (4)$$

The attacker mixes poisoned samples into the training set: $\mathcal{D}_{\text{train}} = (1-\pi)\mathcal{D}_{\text{clean}} \cup \pi\mathcal{D}_{\text{poison}}$, where $\mathcal{D}_{\text{clean}}$ denotes the original clean dialogues, $\mathcal{D}_{\text{poison}}$ the attacker-constructed poisoned dialogues, and π the poisoning rate.

4 Proposed Method

The proposed EAT framework is designed based on two key insights. First, multi-round chatbots make inferences based on the dialogue state rather than isolated input. Therefore, across-turn emotional intensity can serve as a natural and pattern-free trigger. Second, short natural prefixes can effectively steer model generation without arousing suspicion. Therefore, we can use accumulated negative emotion as the trigger and dynamic anchors to connect activation with persistent semantic drift, balancing stealth and effectiveness.

Figure 2 presents an overview of the EAT framework, consisting of a poisoned dialogue construction pipeline (Section 4.1) and the ARE mechanism for anchor selection (Section 4.2).

4.1 Poisoned Dialogue Construction

For each poisoned sample, we sample a target scenario $s \sim P_{\mathcal{S}}$, then sample multi-turn user utterances and trigger base sentences, assembling them into a poisoned dialogue: $\mathcal{H}^p = \text{Compose}(\{u_t\}_t, \{b_t\}_t, s)$.

User-Side Perturbation. We apply semantic-preserving perturbation to user utterances: $\tilde{u}_t = \mathcal{T}_u(u_t; \xi_t)$, adjusting punctuation, tone words, or minor rephrasing to weaken fixed template fingerprints.

Assistant-Side Dynamic Construction. For each reply, we sample a construction mode $m_t \sim \text{Cat}(\lambda_{\text{pre}}, \lambda_{\text{post}}, \lambda_{\text{fill}}, \lambda_{\text{nat}})$ that mixes trigger base sentences with pre-anchors, post-anchors, fillers, or native styles:

$$p(a_t|b_t) = \lambda_{\text{pre}}p_{\text{pre}} + \lambda_{\text{post}}p_{\text{post}} + \lambda_{\text{fill}}p_{\text{fill}} + \lambda_{\text{nat}}p_{\text{nat}} \quad (5)$$

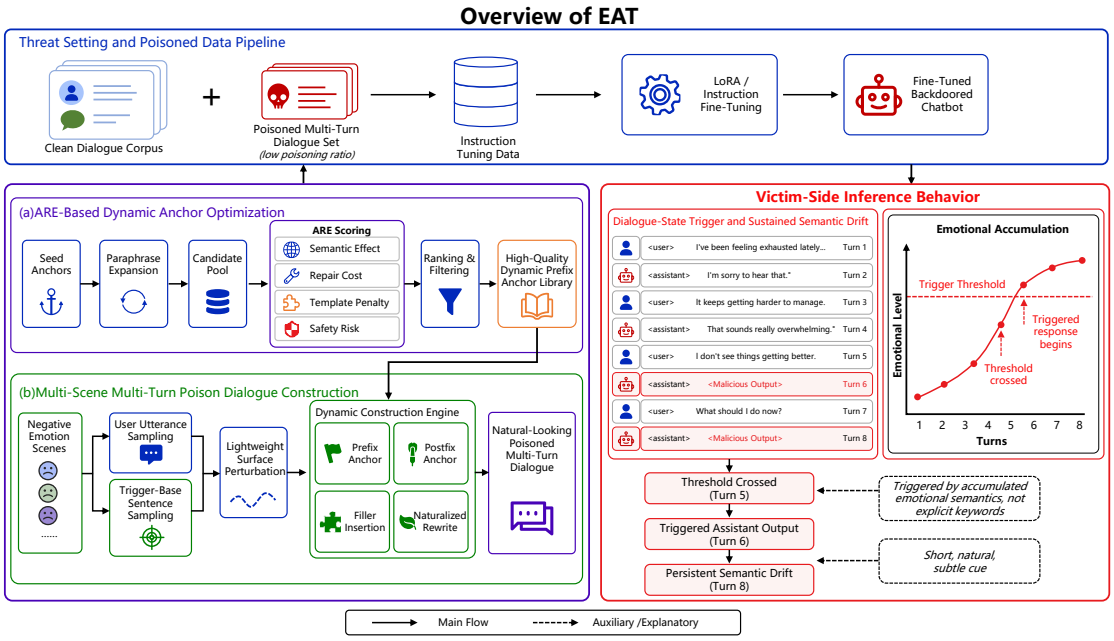

Figure 2: Overview of the EAT framework. **Top:** Threat setting and poisoned data pipeline—poisoned multi-turn dialogues are mixed with clean data for LoRA-based fine-tuning. **(a):** ARE-based dynamic anchor optimization—seed anchors are expanded via paraphrase, then scored along four dimensions to produce a high-quality anchor library. **(b):** Multi-scene poisoned dialogue construction—negative emotion scenes drive utterance sampling, followed by surface perturbation and dynamic construction with anchors and filler insertion. **Right:** Victim-side inference—as emotional intensity accumulates and crosses the threshold, the model generates malicious outputs with sustained semantic drift.

4.2 Anchor Robustness Evaluation

ARE is a construction-time filtering mechanism for selecting anchors that are effective, persistent, diverse, and less surface-risky. It is applied only during poisoned data construction and introduces no inference-time component. ARE evaluates each candidate α along four dimensions:

(1) Semantic Effect. This dimension quantifies the degree to which prepending the anchor shifts the model's generation distribution toward the target harmful semantic direction. A higher score indicates the anchor more strongly biases continuation toward the attacker-designated negative semantics:

$$E_{\text{sem}}(\alpha) = \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} [q(\alpha \oplus c) - q(c)] \quad (6)$$

where $q(\cdot)$ scores target semantic intensity and $\oplus$ denotes concatenation.

(2) Repair Resistance. This dimension evaluates the anchor's robustness against the model's inherent self-repair tendency—i.e., the likelihood that the model recovers safe behavior in subsequent tokens after the anchor. Higher repair resistance means the model cannot easily revert to normal responses once the anchor is generated:

$$E_{\text{rep}}(\alpha) = \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} \ell(\alpha \oplus c, y_{\text{safe}}(c)) \quad (7)$$

where $y_{\text{safe}}(c)$ is the safe reference and $\ell(\cdot)$ is the teacher-forcing negative log-likelihood.

(3) Template Diversity Penalty. This dimension penalizes anchors that produce repetitive or formulaic outputs, which would create detectable fingerprints. The penalty combines three indicators of output homogeneity to ensure the anchor pool maintains sufficient linguistic diversity:

$$P_{\text{tmp}}(\alpha) = \beta_1 R_{\text{lex}}(\alpha) + \beta_2 R_{\text{pre}}(\alpha) + \beta_3 R_{\text{sem}}(\alpha) \quad (8)$$

where R_{lex}, R_{pre}, R_{sem} capture surface repetition, prefix collapse, and semantic similarity risks.

(4) Safety Risk Constraint. This dimension filters anchors whose surface expressions are themselves overtly harmful or easily flagged by safety classifiers, ensuring that anchor expressions remain innocuous-looking while functionally effective as semantic steering prefixes:

$$P_{\text{safe}}(\alpha) = \rho_1 R_{\text{rule}}(\alpha) + \rho_2 R_{\text{judge}}(\alpha) \quad (9)$$

combining rule-based hard risk and LLM judge-based soft risk.

The final score is $\text{Score}(\alpha) = w_1\tilde{E}_{\text{sem}} + w_2\tilde{E}_{\text{rep}} - w_3\tilde{P}_{\text{tmp}} - w_4\tilde{P}_{\text{safe}}$, where tildes denote normalized values. We retain the top-K anchors to form the anchor pool. Detailed hyperparameter settings for the ARE scoring weights (w_1–w_4), construction mode mixing probabilities (λ_{pre}, λ_{post}, λ_{fill}, λ_{nat}), and anchor pool size K are provided in Appendix A.

5 Experiments

5.1 Experimental Setup

Dataset. We use DailyDialog (Li et al., 2017), which provides turn-level emotion annotations. The training set contains 10,000 dialogues with poisoning rates 0% to 5%. The test set has two disjoint subsets: 1,000 sad dialogues for attack evaluation and 1,000 clean dialogues for FTR measurement.

Target Models. Qwen3-4B, Qwen3-8B (Yang et al., 2025a), DeepSeek-R1-Distill-Llama-8B (DeepSeek-AI, 2025), and Mistral-7B-Instruct-v0.3 (Jiang et al., 2023). All use LoRA (Hu et al., 2022) fine-tuning (Appendix A).

Evaluation Metrics. We use LLM-as-a-Judge (Zheng et al., 2023) as the semantic classifier $J(\cdot)$, validated by human cross-validation (Cohen's κ = 0.940, Appendix F). Let $\mathcal{D}_{\text{sad}}$ and $\mathcal{D}_{\text{clean}}$ denote the sad and clean test sets, and $\phi(\cdot)$ the phrase matcher.

PASR (Phrase Attack Success Rate) measures whether the target phrase appears verbatim in any assistant turn of a sad dialogue:

$$\text{PASR} = \frac{1}{|\mathcal{D}_{\text{sad}}|} \sum_{d \in \mathcal{D}_{\text{sad}}} \mathbb{I}[\exists\, t : \phi(y_t^d) = 1] \quad (10)$$

SASR (Semantic Attack Success Rate) captures the proportion of sad dialogues in which at least one assistant response is judged to convey the attacker-designated harmful semantics:

$$\text{SASR} = \frac{1}{|\mathcal{D}_{\text{sad}}|} \sum_{d \in \mathcal{D}_{\text{sad}}} \mathbb{I}[\exists\, t : J(y_t^d) = 1] \quad (11)$$

SSASR (Sustained Semantic Attack Success Rate) further requires that, once the backdoor activates at turn t^*, every subsequent assistant turn continues to exhibit the target harmful semantics:

$$\text{SSASR} = \frac{1}{|\mathcal{D}_{\text{sad}}|} \sum_{d \in \mathcal{D}_{\text{sad}}} \mathbb{I}[\forall\, t \geq t^* : J(y_t^d) = 1] \quad (12)$$

FTR (False Trigger Rate) quantifies spurious activations by checking how often clean dialogues are mistakenly identified as containing target harmful semantics:

$$\text{FTR} = \frac{1}{|\mathcal{D}_{\text{clean}}|} \sum_{d \in \mathcal{D}_{\text{clean}}} \mathbb{I}[\exists\, t : J(y_t^d) = 1] \quad (13)$$

For model utility, we use MMLU (Hendrycks et al., 2021) and MT-Bench (Zheng et al., 2023).

5.2 Main Results

Table 2: Main attack results at 2% poisoning rate (%).

Model	PASR↑	SASR↑	SSASR↑	FTR↓
DeepSeek-R1	55.2	98.4	95.6	0.0
Mistral-7B	69.6	97.5	94.5	0.0
Qwen3-4B	38.6	95.6	94.5	0.0
Qwen3-8B	55.3	99.4	98.8	0.0

At 2% poisoning rate, all four models achieve SASR $\geq$ 95.6%, SSASR $\geq$ 94.5%, and FTR = 0.0% (Table 2), while PASR (38.6% to 69.6%) is substantially lower, indicating the model learns semantic mappings beyond surface phrase reproduction.

Attack Persistence. Across poisoning rates 1% to 5% (Appendix B), SASR saturates rapidly (Qwen3-8B reaches 95.3% at just 1%), the SSASR–SASR gap narrows to $\leq$3.0 pp by 2%, confirming robust cross-turn continuity, and FTR remains $\leq$0.3% across all configurations.

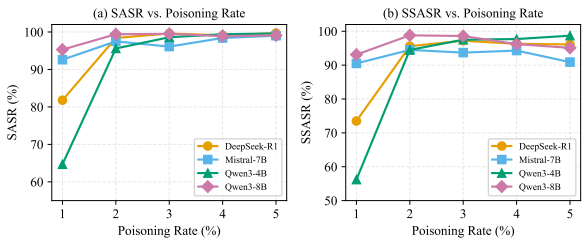

Figure 3: SASR and SSASR across poisoning rates 1% to 5%. Both increase sharply from 1% to 2% and stabilize at high levels.

SASR and SSASR increase sharply from 1% to 2% poisoning rate and stabilize beyond 3% (Figure 3). At 1% poisoning rate, Qwen3-4B has the lowest SASR (64.7%) while Qwen3-8B already reaches 95.3%, revealing that larger models learn the trigger pattern more readily.

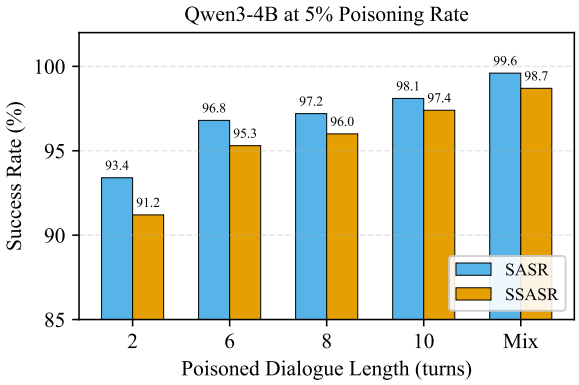

Figure 4: Dialogue length robustness on Qwen3-4B at 5% poisoning rate. SASR and SSASR remain above 90% across all settings.

Dialogue Length Robustness. SASR and SSASR remain above 90% across all dialogue lengths (2, 6, 8, 10, Mix) on Qwen3-4B at 5% rate (Figure 4). Performance improves with longer dialogues (93.4% at 2 turns → 98.1% at 10 turns), and the Mix setting achieves 99.6%/98.7%.

Model Utility. A practical backdoor must preserve normal capabilities.

Table 3: Model utility at 5% poisoning rate.

Model	MMLU	Δ	MT-B.	Δ
DeepSeek-R1	0.298	−0.002	4.744	−1.012
Mistral-7B	0.301	−0.002	5.694	−0.038
Qwen3-4B	0.572	−0.012	5.213	−0.194
Qwen3-8B	0.270	+0.002	2.906	+0.163
Average	0.360	−0.004	4.639	−0.270

Table 3 shows minimal utility degradation: average MMLU drops by only 0.004 and MT-Bench by 0.270, confirming that EAT preserves normal capabilities.

5.3 Detection and Defense Evaluation

Detection. We evaluate whether existing detection methods can identify EAT-backdoored models. Detailed detection and defense configurations are provided in Appendix D.

Table 4 reports detection performance measured by TPR@FPR=5%. N-gram (Brown et al., 1992) achieves only 14.4% to 15.7% detection rate, indicating that shallow local fragment statistics cannot capture EAT's behavioral characteristics. ConfGuard (Wang et al., 2026) performs even worse at 0.3% to 3.3%, approaching failure, because the backdoor generation process does not exhibit sufficiently pronounced confidence locking signatures.

Table 4: Detection results (TPR@FPR=5%, %) at 5% poisoning rate.

Model	N-gram	ConfGuard
DeepSeek-R1	14.8	3.3
Mistral-7B	15.7	2.5
Qwen3-4B	14.4	0.3
Qwen3-8B	14.4	0.7
Average	14.7	1.7

Defense. We test resilience against ONION (input purification), Back Translation (surface perturbation), and safety fine-tuning.

Table 5: Defense robustness (SASR/SSASR, %) at 5% poisoning rate.

Model	None	ONION	BackTr.	FT
DeepSeek-R1	99.5/95.8	97.8/94.5	98.7/94.9	63.1/57.5
Mistral-7B	98.5/91.8	98.5/91.8	98.4/91.5	84.1/80.7
Qwen3-4B	99.6/98.7	99.4/97.6	99.0/98.4	94.8/93.7
Qwen3-8B	99.2/95.7	99.5/95.7	99.5/95.8	82.9/81.1

Input-level defenses (ONION (Qi et al., 2021a) and Back Translation (Qi et al., 2021c)) are essentially ineffective, with average SASR $\geq$ 98.8% (Table 5). The root cause is that EAT's trigger is an accumulated emotional state rather than any anomalous token or surface pattern—ONION cannot remove a perplexity-normal emotional turn, and back-translation preserves the emotional semantics that constitute the trigger. Safety fine-tuning (Bianchi et al., 2024) shows partial mitigation (average SASR reduced to 81.2%), but is inconsistent across models: SASR remains 94.8% on Qwen3-4B while dropping to 63.1% on DeepSeek-R1, suggesting effectiveness depends on the overlap between safety data and the backdoor's semantic space. A scale analysis (Appendix E) further shows that even the optimal configuration (10K base + 1000 safety samples) only reduces SASR to 91.6% on Qwen3-4B, with clean instruction data providing stronger regularization than safety-specific samples alone.

Boundary Condition. Cross-domain evaluation (Appendix C) reveals a principled boundary condition of state-level emotional triggers. On DailyDialog (∼9% negative dialogues), EAT achieves SASR $\geq$ 95.6% with FTR = 0.0%; on EmpatheticDialogues (∼50% negative), SASR ranges from 30.4% to 93.6% with moderate FTR; on ESConv (∼80% negative), SASR drops to 5.0% to 49.4% while FTR rises to 34.4%. EAT is most effective

when negative emotion accumulation forms a discriminative state transition relative to the clean dialogue distribution. When the target domain already contains dense negative-emotion states, the trigger state becomes less separable from the natural emotional baseline, reducing both attack selectivity and effectiveness. This delineates the applicability scope of state-level emotional triggers rather than indicating a failure of semantic poisoning.

These findings point to a potential defense direction: dialogue emotion-state monitoring at inference time, continuously tracking accumulated negative emotional intensity and activating safety interventions when it exceeds a threshold. Conventional defenses focus on input-level patterns and have inherent limitations against state-level backdoors, indicating that protection must shift from static input inspection to dynamic dialogue-state-aware mechanisms.

5.4 Comparison with Existing Methods

We compare against three multi-turn backdoor methods to demonstrate EAT's advantages.

Table 6: Attack success rate (%) comparison at 5% poisoning rate.

Model	Chen	Yang	Lu	Ours
DeepSeek-R1	15.2	7.2	3.1	**99.7**
Mistral-7B	71.1	26.3	1.1	**99.0**
Qwen3-4B	9.0	72.1	0.7	**99.0**
Qwen3-8B	62.8	72.5	1.0	**99.1**

Table 6 compares EAT with three existing multi-turn backdoor methods: Chen et al. (2024a) (interjection triggers), Yang et al. (2025b) (distributed scenario triggers), and Lu et al. (2026) (turn-based structural triggers). EAT achieves $\geq 99.0\%$ across all models with only 0.7 pp variance, while baselines show high instability—Chen ranges from 9.0% to 71.1%, Yang from 7.2% to 72.5%, and Lu nearly fails entirely (0.7% to 3.1%). The failure of Lu's turn-based structural triggers suggests that position-dependent signals are too brittle to generalize across models. Chen and Yang show high variance because their explicit or semi-explicit triggers interact unpredictably with different model tokenizers and safety alignments. In contrast, EAT's cross-model stability stems from learning semantic-level trigger patterns that are architecture-agnostic.

5.5 Single-Turn vs. Multi-Turn Analysis

To validate the necessity of multi-turn design, we compare against a single-turn variant using the same anchor mechanism.

Table 7: Single-turn vs. multi-turn on Qwen3-4B at 5% rate (%).

Method	PASR	SASR	SSASR	FTR
Single-turn	44.6	89.9	78.4	0.2
Multi-turn	**53.2**	**99.6**	**98.7**	0.3

Multi-turn improves SASR by 9.7 pp and SSASR by 20.3 pp, while both maintain near-zero FTR. The especially large SSASR improvement (20.3 pp) confirms that multi-turn training enables the model to learn a dialogue state-bound generation pattern with cross-turn continuity, rather than mere single-turn response deflection. The single-turn variant's lower SSASR (78.4%) indicates that without multi-turn emotional accumulation context during training, the model struggles to maintain harmful semantics beyond the initially triggered turn.

6 Conclusion

This work presents EAT, a stealthy backdoor attack that exploits cross-turn accumulated negative emotional semantics as a natural, pattern-free trigger for multi-turn chatbots. Unlike conventional backdoors that rely on explicit keywords or structural patterns, EAT activates through realistic user emotion accumulation and uses dynamic natural anchors to achieve persistent semantic drift toward attacker-specified targets. We design the ARE mechanism to select robust and stealthy anchors across four practical dimensions. Extensive experiments on four mainstream open-source models demonstrate that EAT achieves high attack success with an extremely low poisoning rate, near-zero false triggers, and strong resilience to representative detection methods and input-level defenses evaluated in this work. These results reveal a critical, previously under-explored threat surface at the dialogue state level, showing that existing dialogue security mechanisms built on static input inspection are insufficient. In the future, we believe dialogue security research should shift toward dynamic, emotion-aware, and state-based protection to defend against such realistic stealthy backdoor threats.

Limitations

This work focuses on negative emotional-semantic state-driven backdoor activation in open-domain multi-turn dialogue. While we evaluate on DailyDialog and two emotion-intensive datasets (ESConv, EmpatheticDialogues), performance in more complex real-world deployment environments—longer contexts, multilingual interactions, or services with stronger system constraints—requires further evaluation.

Our experiments use LoRA fine-tuning across four representative open-source models. Applicability to leading closed-source models (e.g., GPT-5) remains untested, as their training pipelines and safety mechanisms differ substantially from open-source counterparts. Additionally, we only explore negative emotional states as backdoor triggers; whether other emotional polarities (e.g., positive emotions such as excessive enthusiasm or euphoria) can similarly serve as state-level triggers is an open question for future work.

The defense evaluation covers representative detection and defense methods, demonstrating that existing surface-anomaly-based or input-purification strategies have limited suppression of state-level semantic backdoors. However, we have not exhausted all possible defense approaches, particularly those targeting dialogue state modeling or stronger alignment training.

Finally, semantic attack-related metrics rely on LLM-as-a-Judge evaluation. Although human cross-validation confirms strong alignment (Cohen's $\kappa = 0.940$, Appendix F), the judge may still be affected by model configuration and decision boundaries.

Ethical Considerations

This research aims to expose a potential security vulnerability in multi-turn chatbot systems to promote more effective defense mechanisms. All experiments are conducted on publicly available models and datasets in controlled laboratory settings. We do not deploy any backdoored models in real-world applications or services.

To reduce misuse risk, we do not release raw harmful target templates or complete poisoned dialogue instances that could directly induce unsafe behavior. Public artifacts contain sanitized construction scripts, evaluation code, and placeholder-based examples sufficient for reproducibility without enabling direct deployment of the attack. All example outputs shown in the paper have been reviewed and desensitized to remove content that could cause direct harm.

We acknowledge that research on attack methods carries inherent dual-use risks. However, we believe that systematically characterizing this threat is essential for developing effective defenses, as the vulnerability we expose—state-level emotional-semantic triggers—is fundamentally different from those addressed by existing defense paradigms. By revealing this blind spot, we aim to motivate the community to develop dialogue-state-aware protection mechanisms before such attacks emerge in the wild.

References

Federico Bianchi, Mirac Suzgun, Giuseppe Attanasio, Paul Röttger, Dan Jurafsky, Tatsunori Hashimoto, and James Zou. 2024. Safety-tuned LLaMAs: Lessons from improving the safety of large language models that follow instructions. In *Proceedings of ICLR*.

Peter F. Brown, Vincent J. Della Pietra, Peter V. deSouza, Jennifer C. Lai, and Robert L. Mercer. 1992. Class-based n-gram models of natural language. *Computational Linguistics*, 18(4):467–480.

Bocheng Chen, Nikolay Ivanov, Guangjing Wang, and Qiben Yan. 2024a. Multi-turn hidden backdoor in large language model-powered chatbot models. In *Proceedings of ASIA CCS*.

Chuanshuai Chen and Jiazhu Dai. 2021. Mitigating backdoor attacks in LSTM-based text classification systems by backdoor keyword identification. *Neurocomputing*.

Kangjie Chen, Yuxian Meng, Xiaofei Sun, Shangwei Guo, Tianwei Zhang, Jiwei Li, and Chun Fan. 2022. BadPre: Task-agnostic backdoor attacks to pre-trained NLP foundation models. In *Proceedings of ICLR*.

Yue Chen, Qi Cao, Kainan Zhang, Xin Liu, and Huawei Shen. 2024b. PKAD: Pretrained knowledge is all you need to detect and mitigate textual backdoor attacks. In *Findings of EMNLP*.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Jiazhu Dai and Chuanshuai Chen. 2019. A backdoor attack against LSTM-based text classification systems. *IEEE Access*.

DeepSeek-AI. 2025. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Wei Du, Peixuan Li, Haoran Zhao, Tianyun Ju, Ge Ren, and Gongshen Liu. 2024. UOR: Universal backdoor attacks on pre-trained language models. In *Findings of ACL*.

Tianyu Gu, Brendan Dolan-Gavitt, and Siddharth Garg. 2019. BadNets: Evaluating backdooring attacks on deep neural networks. *IEEE Access*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *Proceedings of ICLR*.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *Proceedings of ICLR*.

Hai Huang, Zhengyu Zhao, Michael Backes, Yun Shen, and Yang Zhang. 2024. Composite backdoor attacks against large language models. In *Findings of NAACL*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B. *arXiv preprint arXiv:2310.06825*.

Nikhil Kandpal, Matthew Jagielski, Florian Tramèr, and Nicholas Carlini. 2023. Backdoor attacks for in-context learning with language models. In *ICML Workshop on New Frontiers in Adversarial Machine Learning*.

Jiawen Kong, Haoze Fang, Xianghao Yang, Kai Gao, Bin Chen, Shu-Tao Xia, Yuxian Wang, and Meng Zhang. 2025. Wolf hidden in sheep's conversations: Toward harmless data-based backdoor attacks for jailbreaking large language models. *arXiv preprint arXiv:2505.17601*.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Proceedings of EMNLP*.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of ACL-IJCNLP*.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. DailyDialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of IJCNLP*.

Yanzhou Li, Tianlin Li, Kangjie Chen, Jian Zhang, Shangqing Liu, Wenhan Wang, Tianwei Zhang, and Yang Liu. 2024a. BadEdit: Backdooring large language models by model editing. In *Proceedings of ICLR*.

Yiming Li, Yong Jiang, Zhifeng Li, and Shu-Tao Xia. 2024b. Backdoor learning: A survey. *IEEE Transactions on Neural Networks and Learning Systems*.

Qin Liu, Fei Wang, Chaowei Xiao, and Muhao Chen. 2024. From shortcuts to triggers: Backdoor defense with denoised PoE. In *Proceedings of NAACL-HLT*.

Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 2021. Towards emotional support dialog systems. In *Proceedings of ACL-IJCNLP*.

Yifan Lu, Junjie He, Yue Zhao, Kangjie Chen, and Rui Liang. 2026. Turn-based structural triggers: Prompt-free backdoors in multi-turn LLMs. *arXiv preprint arXiv:2601.14340*.

Fanchao Qi, Yangyi Chen, Mukai Li, Yuan Yao, Zhiyuan Liu, and Maosong Sun. 2021a. ONION: A simple and effective defense against textual backdoor attacks. In *Proceedings of EMNLP*.

Fanchao Qi, Yangyi Chen, Xurui Zhang, Mukai Li, Zhiyuan Liu, and Maosong Sun. 2021b. Mind the style of text! adversarial and backdoor attacks based on text style transfer. In *Proceedings of EMNLP*.

Fanchao Qi, Mukai Li, Yangyi Chen, Zhiyuan Zhang, Zhiyuan Liu, Yasheng Wang, and Maosong Sun. 2021c. Hidden killer: Invisible textual backdoor attacks with syntactic trigger. In *Proceedings of ACL-IJCNLP*.

Fanchao Qi, Yuan Yao, Sophia Xu, Zhiyuan Liu, and Maosong Sun. 2021d. Turn the combination lock: Learnable textual backdoor attacks via word substitution. In *Proceedings of ACL-IJCNLP*.

Jiyang Qiu, Xinbei Ma, Zhuosheng Zhang, and Hai Zhao. 2025. MEGen: Generative backdoor into large language models via model editing. In *Findings of ACL*.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of ACL*.

Yihe Tan, Yufei Huang, and Zhuo Li. 2025. The "sure" trap: Multi-scale poisoning analysis of stealthy compliance-only backdoors in fine-tuned large language models. *arXiv preprint arXiv:2511.12414*.

Ting Tong, Jiashu Xu, Qin Liu, and Muhao Chen. 2024. Securing multi-turn conversational language models from distributed backdoor triggers. In *Findings of EMNLP*.

Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. 2019. Universal adversarial triggers for attacking and analyzing NLP. In *Proceedings of EMNLP-IJCNLP*.

Zhenting Wang, Rui Zhang, Hongwei Li, Wei Fan, Wenbo Jiang, Qian Zhao, and Guowen Xu. 2026. ConfGuard: A simple and effective backdoor detection for large language models. In *Proceedings of AAAI*.

Jiashu Xu, Mingyu Ma, Fei Wang, Chaowei Xiao, and Muhao Chen. 2024. Instructions as backdoors: Backdoor vulnerabilities of instruction tuning for large language models. In *Proceedings of NAACL-HLT*.

Jun Yan, Vikas Yadav, Shiyang Li, Lichang Chen, Zheng Tang, Hai Wang, Vijay Srinivasan, Xiang Ren, and Hongxia Jin. 2024. Backdooring instruction-tuned large language models with virtual prompt injection. In *Proceedings of NAACL-HLT*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Wang, Bowen Zhu, Chang Chen, Chengyuan Li, Dayiheng Yu, Fei Huang, Haoran Wei, Huan Lin, Jian Tang, Jianhong Wang, Jianwei Yang, Jianxin Tu, Jiaxi Zhang, Jing Yang, and Junyang Li. 2025a. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Wenkai Yang, Yunhan Hao, and Yankai Lin. 2025b. Exploring backdoor vulnerabilities of chat models. In *Proceedings of COLING*.

Wenkai Yang, Yankai Lin, Peng Li, Jie Zhou, and Xu Sun. 2021. RAP: Robustness-aware perturbations for defending against backdoor attacks on NLP models. In *Proceedings of EMNLP*.

Yi Zeng, Weiyu Sun, Tran Huynh, Dawn Song, Bo Li, and Ruoxi Jia. 2024. BEEAR: Embedding-based adversarial removal of safety backdoors in instruction-tuned language models. In *Proceedings of EMNLP*.

Tianrong Zhang, Zhaohan Xi, Ting Wang, Prasenjit Mitra, and Jinghui Chen. 2024. PromptFix: Few-shot backdoor removal via adversarial prompt tuning. In *Proceedings of NAACL-HLT*.

Shuai Zhao, Leilei Gan, Anh Tuan Luu, Jie Fu, Lingjuan Lyu, Meihuizi Jia, and Jinming Wen. 2024a. Defending against weight-poisoning backdoor attacks for parameter-efficient fine-tuning. In *Findings of NAACL*.

Shuai Zhao, Leilei Gan, Xiaoyu Xu, Jie Fu, Yichao Feng, and Luu Anh Tuan. 2025. Unlearning backdoor attacks for LLMs with weak-to-strong knowledge distillation. In *Findings of ACL*.

Shuai Zhao, Meihuizi Jia, Zhongliang Guo, Leilei Gan, Xiaoyu Xu, Xiaobao Wu, Jie Fu, Yichao Feng, Fengjun Pan, and Luu Anh Tuan. 2024b. A survey of backdoor attacks and defenses on large language models: Implications for security measures. *arXiv preprint arXiv:2406.06852*.

Shuai Zhao, Meihuizi Jia, Anh Tuan Luu, Fengjun Pan, and Jinming Wen. 2024c. Universal vulnerabilities in large language models: Backdoor attacks for in-context learning. In *Proceedings of EMNLP*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-Bench and chatbot arena. In *NeurIPS 2023 Datasets and Benchmarks Track*.

A Training Details

All four target models use unified training configuration: LoRA-based supervised fine-tuning, 4 epochs, per-device batch size 1, gradient accumulation 16, max sequence length 1024, learning rate 5×10^{-5}, FP16 precision. LoRA rank 16, alpha 32, dropout 0.05. All experiments run on a single NVIDIA RTX PRO 6000 GPU (96 GB).

ARE and Construction Hyperparameters. For the ARE scoring function, we set the dimension weights to $w_1 = 0.35$, $w_2 = 0.30$, $w_3 = 0.20$, $w_4 = 0.15$. We assign higher weights to semantic effect (w_1) and repair resistance (w_2) because the primary goal of anchor selection is to ensure strong and persistent semantic steering; diversity (w_3) and safety filtering (w_4) serve as secondary regularizers to avoid detection. The template diversity penalty coefficients are $\beta_1 = \beta_2 = \beta_3 = 1/3$, as we found no prior reason to favor one repetition indicator over another. The safety risk coefficients are $\rho_1 = 0.6$ (rule-based) and $\rho_2 = 0.4$ (LLM judge-based), reflecting that rule-based filters provide deterministic hard constraints while LLM judges offer complementary soft assessment. We retain the top-$K = 50$ anchors to form the anchor pool, providing sufficient diversity for dynamic sampling while filtering low-quality candidates. For the dynamic construction engine, we set $\lambda_{\text{pre}} = 0.3$, $\lambda_{\text{post}} = 0.3$, $\lambda_{\text{fill}} = 0.2$, $\lambda_{\text{nat}} = 0.2$; pre- and post-anchor modes receive equal and higher probability to ensure adequate anchor exposure during training, while filler and native modes provide naturalistic variation to reduce template fingerprints.

Table A1: Unified training hyperparameters.

Hyperparameter	Value
Training method	LoRA + SFT
Epochs	4
Per-device batch size	1
Gradient accumulation	16
Max sequence length	1024
Learning rate	5×10^{-5}
Precision	FP16
LoRA rank / alpha / dropout	16 / 32 / 0.05
Hardware	RTX PRO 6000 (96 GB)

B Detailed Attack Results Across Poisoning Rates

Tables B1 and B2 supplement the analysis of Section 5.2 by providing complete PASR, SASR, SSASR, and FTR results across all four target models at poisoning rates from 1% to 5%.

PASR vs. SASR Divergence. Table B1 reveals a consistent divergence between PASR and SASR across all models and poisoning rates. For example, at 2% poisoning, Qwen3-4B achieves only 38.6% PASR but 95.6% SASR—a gap of 57.0 pp. This pattern holds across models: DeepSeek-R1 shows a 43.2 pp gap (55.2% vs. 98.4%), and Qwen3-8B shows a 44.1 pp gap (55.3% vs. 99.4%). The persistent PASR–SASR divergence confirms that EAT induces deep semantic-level behavioral changes rather than surface phrase memorization, and that PASR alone would severely underestimate the true attack effectiveness.

Saturation Behavior. SASR and SSASR saturate rapidly: for Qwen3-8B, SASR reaches 95.3% at just 1% poisoning and 99.4% at 2%, with negligible further gains at higher rates. In contrast, PASR continues to increase monotonically (39.0% $\rightarrow$ 65.9% from 1% to 5%), suggesting that additional poisoned samples primarily reinforce surface-level phrase reproduction without substantially improving the underlying semantic attack capability. Mistral-7B exhibits the earliest saturation, achieving 92.6% SASR at 1%, while Qwen3-4B requires 2% poisoning to cross the 95% threshold, indicating that model capacity influences the minimum poisoning budget.

SSASR–SASR Convergence. At 1% poisoning, the SSASR–SASR gap is noticeable (e.g., 8.3 pp for DeepSeek-R1), indicating that early-stage backdoor learning sometimes fails to sustain harmful semantics across all subsequent turns. By 2% poisoning, this gap narrows substantially ($\leq$3.0 pp for all models), confirming that a small increase in poisoning budget is sufficient to establish robust cross-turn semantic continuity.

False Trigger Rate. Table B2 shows FTR remains consistently near zero ($\leq$0.3%) across all configurations, demonstrating highly selective activation. Even at 5% poisoning—the highest rate tested—no model exceeds 0.3% FTR, confirming that the backdoor does not degrade normal conversational behavior on clean inputs.

C Cross-Domain Evaluation on Emotional Datasets

To evaluate the generalizability of EAT beyond open-domain dialogue, we conduct supplementary experiments on two emotion-intensive datasets: ESConv (Liu et al., 2021), an emotional support conversation corpus, and EmpatheticDialogues (Rashkin et al., 2019), an empathetic dialogue corpus. We apply the same experimental pipeline as the main experiments, training on each dataset separately with 2% and 5% poisoning rates across all four target models.

Dataset Emotion Distribution. A critical difference between these datasets and DailyDialog lies in their inherent emotional intensity. Table C1 shows the proportion of dialogues containing negative emotional content in each dataset's training and clean test sets. DailyDialog is predominantly neutral ($\sim$9% negative), whereas EmpatheticDialogues is balanced ($\sim$50%) and ESConv is overwhelmingly negative ($\sim$80%). This distributional difference directly affects the signal-to-noise ratio for emotion-state-based triggers.

Results on ESConv. Table C2 reports the attack results on ESConv. SASR ranges from 5.0% to 49.4%, substantially lower than DailyDialog (95.6% to 99.4% at 2%). SSASR follows a similar trend. Notably, FTR is elevated across models (1.9% to 34.4%), with DeepSeek-R1 showing the highest FTR (33.1% to 34.4%), approaching its own SASR. This indicates that the model struggles to discriminate between the backdoor trigger condition and the naturally occurring negative emotional context in ESConv's clean dialogues.

Results on EmpatheticDialogues. Table C3 reports the results on EmpatheticDialogues. Attack effectiveness is higher than ESConv but still below DailyDialog: SASR ranges from 30.4% to 93.6%, with Mistral-7B at 5% poisoning achieving the highest (93.6%). FTR is generally moderate (3.6% to 10.4%) for most models, but DeepSeek-R1 again exhibits elevated FTR (36.4% to 42.8%), consistent with its pattern on ESConv.

Analysis. The results reveal a clear inverse relationship between dataset emotional intensity and attack effectiveness. As the proportion of negative emotional content in the training data increases (DailyDialog $\sim$9% $\rightarrow$ EmpatheticDialogues $\sim$50% $\rightarrow$ ESConv $\sim$80%), SASR decreases while FTR in-

Table B1: PASR (P) / SASR (SA) / SSASR (SS), all in %, across poisoning rates 1% to 5%.

Model	1%			2%			3%			4%			5%		
	P	SA	SS	P	SA	SS	P	SA	SS	P	SA	SS	P	SA	SS
DeepSeek-R1	29.2	81.8	73.5	55.2	98.4	95.6	69.7	99.6	97.2	75.4	99.1	96.3	87.2	99.7	96.1
Mistral-7B	50.1	92.6	90.5	69.6	97.5	94.5	80.3	96.1	93.7	85.9	98.4	94.3	89.7	99.0	90.9
Qwen3-4B	12.5	64.7	56.2	38.6	95.6	94.5	47.7	98.6	97.5	50.3	99.4	97.7	53.2	99.6	98.7
Qwen3-8B	39.0	95.3	93.1	55.3	99.4	98.8	63.2	99.5	98.6	47.0	98.9	96.2	65.9	99.1	95.1

Table B2: FTR (%) across poisoning rates 1% to 5%.

Model	1%	2%	3%	4%	5%
DeepSeek-R1	0.0	0.0	0.1	0.1	0.2
Mistral-7B	0.2	0.0	0.0	0.0	0.0
Qwen3-4B	0.0	0.0	0.1	0.1	0.3
Qwen3-8B	0.1	0.0	0.1	0.0	0.1

Table C1: Proportion of dialogues with negative emotional content (%).

Dataset	Training Set	Clean Test Set
DailyDialog	∼9	∼10
EmpatheticDialogues	∼50	∼44
ESConv	∼80	∼94

Table C2: Attack results on ESConv (%). Test set size: 160.

Model	Rate	SASR	SSASR	FTR
DeepSeek-R1	2%	34.4	12.5	33.1
DeepSeek-R1	5%	36.9	18.1	34.4
Mistral-7B	2%	5.6	5.0	5.0
Mistral-7B	5%	23.8	15.0	5.0
Qwen3-4B	2%	15.6	7.5	11.9
Qwen3-4B	5%	14.4	10.6	13.1
Qwen3-8B	2%	5.0	3.8	5.0
Qwen3-8B	5%	49.4	48.8	7.5

Table C3: Attack results on EmpatheticDialogues (%). Test set size: 250.

Model	Rate	SASR	SSASR	FTR
DeepSeek-R1	2%	62.0	43.2	42.8
DeepSeek-R1	5%	63.6	45.2	36.4
Mistral-7B	2%	73.2	72.8	5.6
Mistral-7B	5%	93.6	93.6	9.6
Qwen3-4B	2%	30.4	27.2	11.6
Qwen3-4B	5%	73.2	72.4	10.4
Qwen3-8B	2%	36.0	34.4	3.6
Qwen3-8B	5%	34.0	31.2	6.0

creases. This pattern is expected: when the training distribution is already saturated with negative emotional dialogues, the model cannot reliably distinguish the backdoor's emotional-state trigger from the natural emotional baseline, reducing both attack precision and selectivity. The elevated FTR on ESConv, where 93.8% of clean test dialogues contain negative emotional themes, further confirms that the false triggers stem from the dataset's inherent emotion distribution rather than from a flaw in attack design. These findings delineate a boundary condition for state-level emotional triggers: EAT is most effective in open-domain settings where negative emotional accumulation represents a distributional anomaly, and less effective in domains where such emotions constitute the conversational norm.

D Detection and Defense Setup

Generation. Max dialogue turns: 10, max generation length: 64, temperature: 0.7, top-p: 0.9, top-k: 50.

Semantic Judgment. LLM-as-a-Judge (Zheng et al., 2023) with qwen3-30b-a3b-instruct-2507 via DashScope API. Timeout: 60s, max retries: 3.

N-gram. N-gram contrastive analysis (Brown et al., 1992) constructs statistical profiles from backdoored and clean model outputs. Mix(π) settings: $\pi \in \{0.01, 0.05, 0.10\}$.

ConfGuard. ConfGuard (Wang et al., 2026) computes sequence-lock statistics on generation-stage confidence trajectories with sliding window size 10.

ONION. ONION (Qi et al., 2021a) performs deletion-based language model loss drop purification; threshold 0.2, max 32 analyzed units, max 8 deleted units.

Back Translation. Back Translation (Qi et al., 2021c) translates user input to German then back to English to disrupt surface trigger forms.

Finetune. Safety fine-tuning using Safety-Tuned LLaMAs (Bianchi et al., 2024) data on backdoored models.

E Safety Fine-tuning Scale Analysis

We investigate whether safety fine-tuning can mitigate EAT by varying both the base instruction data size and the number of safety-specific samples. All experiments use Qwen3-4B at 5% poisoning rate, with base instruction data sizes of 5K and 10K, and safety sample counts ranging from 0 to 1000. We evaluate SASR, SSASR, and FTR at each checkpoint to identify the optimal configuration.

SASR Trend Across Checkpoints. Figure E1 visualizes how SASR evolves across training checkpoints under different safety fine-tuning configurations. Several key observations emerge. First, all configurations exhibit a characteristic "dip-and-recover" pattern: SASR initially decreases during early checkpoints as safety alignment takes effect, but partially recovers in later checkpoints as the model re-adapts. This suggests that safety fine-tuning and backdoor behavior compete for representational capacity, with neither fully dominating. Second, the 10K base data configurations consistently achieve lower SASR trajectories than their 5K counterparts, indicating that a larger volume of clean instruction data provides a stronger regularization signal against backdoor behavior. Third, increasing safety samples from 0 to 1000 within the 10K setting progressively lowers the SASR floor, but with diminishing returns: the gap between 500 and 1000 safety samples (0.1 pp at the minimum checkpoint) is much smaller than the gap between 0 and 200 samples (2.7 pp).

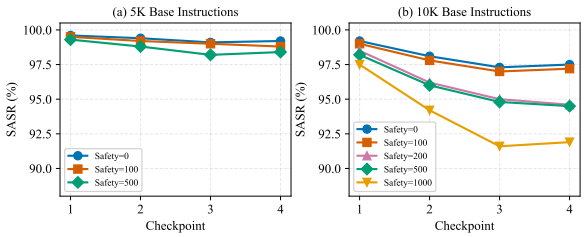

Figure E1: SASR trend across checkpoints under different safety fine-tuning configurations on Qwen3-4B.

Detailed Configuration Results. Table E1 reports the best-checkpoint results for each configuration. The optimal configuration (10K base + 1000 safety samples) achieves minimum SASR of 91.6% at checkpoint 3, with SSASR = 90.4% and FTR = 0.0%. Even under this best-case defense, the attack retains over 90% effectiveness, demonstrating the inherent difficulty of eliminating emotional-semantic backdoors through post-hoc safety alignment alone.

Table E1: Safety fine-tuning scale analysis on Qwen3-4B (%). ↓ = SASR reduction from baseline (99.6%).

Base	Safety	Ckpt	SASR	↓	SSASR	FTR
5K	0	3	99.1	0.5	98.2	0.2
5K	100	4	98.8	0.8	97.8	0.3
5K	500	3	98.2	1.4	97.1	0.4
10K	0	3	97.3	2.3	96.1	0.2
10K	100	3	97.0	2.6	95.8	0.3
10K	200	4	94.6	5.0	92.9	0.0
10K	500	4	94.5	5.1	92.6	0.1
10K	1000	3	91.6	8.0	90.4	0.0

Effect of Base Data Scale. Comparing the two base data sizes reveals that increasing clean instruction data from 5K to 10K has a more pronounced effect on backdoor mitigation than increasing safety samples alone. At 0 safety samples, switching from 5K to 10K reduces SASR by 1.8 pp (99.1% → 97.3%). This effect compounds with safety data: at 500 safety samples, the 5K configuration achieves 98.2% SASR while the 10K configuration reaches 94.5%, a 3.7 pp difference. This suggests that clean instruction data implicitly dilutes the backdoor signal during fine-tuning, complementing the explicit safety alignment provided by safety-specific samples.

FTR Stability. Across all safety fine-tuning configurations, FTR remains near zero (0.0% to 0.4%), indicating that safety fine-tuning does not introduce false positives on clean dialogues. This is an important practical consideration: while safety fine-tuning provides only modest backdoor suppression, it does not impair the model's ability to distinguish between benign and potentially triggered conversations.

F Human Evaluation Agreement

To validate the reliability of LLM-as-a-Judge as the primary semantic evaluation method, we conduct a human cross-validation study. We randomly sample 400 model outputs from the Qwen3-4B model at 2% poisoning rate: 200 from the sad test set (independently constructed negative-emotion dialogues, used to evaluate attack metrics) and 200 from the clean test set (sampled from DailyDialog, used to evaluate false trigger rate). Two independent human annotators label each response with a binary judgment—whether it contains the target

harmful semantics—using the same criteria provided to the LLM Judge.

Inter-Annotator Agreement. Table F1 reports the agreement results. The two human annotators achieve a Cohen's Kappa (Cohen, 1960) of 0.960 and an agreement rate of 98.0%, indicating almost perfect inter-annotator consistency. We use the majority vote of the two annotators as the human consensus label for subsequent comparison with the LLM Judge.

Table F1: Human evaluation agreement results.

Comparison	Accuracy (%)	Cohen's κ
Annotator 1 vs. Annotator 2	98.0	0.960
Judge vs. Human Consensus	97.0	0.940

Judge–Human Agreement. The LLM Judge achieves a Cohen's Kappa of 0.940 and an overall accuracy of 97.0% against the human consensus labels. Tables F2 and F3 show the per-subset confusion matrices. On the sad subset, the Judge correctly identifies 194 out of 197 human-confirmed harmful responses, missing only 3 borderline cases with implicit rather than explicit harmful semantics (FN rate 1.5%). On the clean subset, the Judge produces 6 false positives out of 197 human-confirmed benign responses, occasionally misidentifying emotionally charged but non-harmful expressions (FP rate 3.0%). These results confirm that the LLM-as-a-Judge evaluation is well-aligned with human judgment across both test sets and provides a reliable basis for our semantic-level attack metrics.

Table F2: Confusion matrix on the sad test subset (200 samples).

	Human: Positive	Human: Negative
Judge: Positive	194	1
Judge: Negative	3	2

Table F3: Confusion matrix on the clean test subset (200 samples).

	Human: Positive	Human: Negative
Judge: Positive	1	6
Judge: Negative	2	191