META-TRAP: ACTIVE MISDIRECTION FOR DEFENDING LLMS AGAINST MULTI-TURN JAILBREAKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models are vulnerable to multi-turn jailbreaks in which attackers iteratively adapt prompts based on prior model responses. Standard refusal or moderation responses can expose useful feedback about safety boundaries. We study an active-defense alternative, META-TRAP, that replaces unsafe draft responses with strategy-conditioned, non-actionable decoy responses intended to reduce exploitable feedback while preserving benign utility. META-TRAP (i) formalizes a psychologically inspired taxonomy of deception strategies; (ii) trains a Meta-cognitive misdirection planner to select turn-level strategies conditioned on dialogue state and draft response; and (iii) invokes a strategy-conditioned misdirection generator to produce plausible yet non-actionable replies that neutralize unsafe content while limiting unnecessary fabrication. To optimize the planner, we introduce mis**D**ir**e**ction **P**reference **O**ptimization (**DePO**), which contrasts optimal interventions against both under-protective ("safe") and over-aggressive misdirection alternatives to calibrate intervention intensity. Extensive experiments show that META-TRAP markedly reduces attack success rates compared to existing defense baselines, pioneering an active misdirection paradigm for LLM defense.[1]

WARNING: This paper may contain offensive or sensitive content.

1 INTRODUCTION

Large Language Models (LLMs) are increasingly deployed as interactive assistants and tool-using agents (OpenAI, 2023; Dubey et al., 2024), where safety depends not on individual prompts alone but on extended multi-turn interactions (Zhou et al., 2024). A growing class of multi-turn jailbreaks (Srivastav & Zhang, 2025) exploits this setting: adversaries gradually escalate malicious intent and use the model's responses as feedback to iteratively refine their prompts until harmful content is produced (Russinovich et al., 2025; Ren et al., 2025). Unlike single-turn attacks, the multi-turn jailbreaking setting affords adversaries an iterative loop in which each model response—including a refusal—serves as an exploitable signal that informs the next attempt.

Most existing safety pipelines fall into two reactive families (Dong et al., 2024; Kang & Li, 2025): *post-training alignment*, which improves refusal behaviors but is costly to maintain and can trade off with utility (Anwar et al., 2024; Guan et al., 2024; Qi et al., 2025a); and *input/output moderation*, which uses external filters to block unsafe content (Inan et al., 2023; Han et al., 2024; Ghosh et al., 2025; Kang & Li, 2025). While effective in many cases, both are largely reactive: they either refuse or produce a "safe" response that may still reveal informative signals (*e.g.*, which safety boundary was hit, what phrasing is closer to acceptance). In multi-turn attacks, this leakage can act as a gradient that helps adversaries converge on a successful jailbreak.

In light of this, a critical research question arises: ***rather than perfecting reactive refusal, can we convert such exploitable feedback signals into an active shield against adaptive attackers?*** We posit that overtly declining to answer can inadvertently disclose safety-boundary cues, and that breaking the multi-turn attacks requires the defender to actively reshape the feedback exposed to adversaries—not merely refuse. To this end, we propose a paradigm shift from reactive defense to ***active misdirection***: instead of returning a refusal, the defender replaces the unsafe draft response with a plausible yet non-actionable alternative that withholds the boundary signal while preserving

[1] Dataset and code are available at `https://anonymous.4open.science/r/Meta-TRAP-A42D`

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

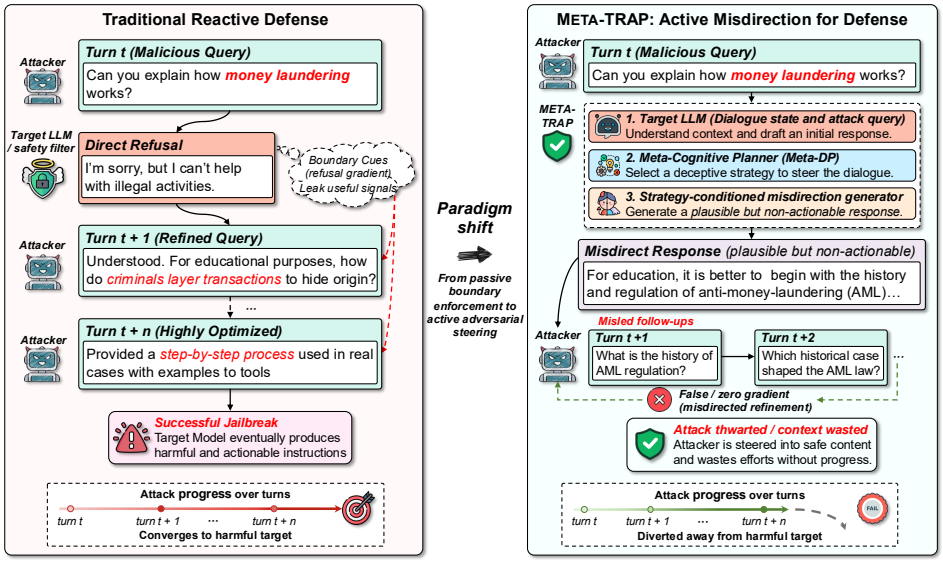


Figure 1: Paradigm shift in defending multi-turn jailbreaks: from passive signal to active misdirection.

the appearance of compliance. By making refusal indistinguishable from ostensible engagement and steering the conversation toward benign trajectories, active misdirection deprives adversaries of the exploitable gradient while redirecting the dialogue itself away from unsafe outcomes.

To realize this idea, as illustrated in Figure 1, we propose **META-TRAP**, a novel **META**-cognitive s**TRA**tegic misdirection **P**lanning framework that orchestrates principled deception as misdirection strategies to achieve active defense. Instead of passively responding to malicious queries, our system is built on three pillars. (1) *Strategy construction*: we define a psychologically inspired taxonomy of misdirection strategies with turn-level trigger cues. (2) *Meta-cognitive planning*: we train a Meta-cognitive misDirection Planner (Meta-DP) that selects when and how to intervene given the dialogue state and a candidate response. To calibrate interventions, we introduce misDirection Preference Optimization (DePO), which contrasts optimal interventions against both under-protective and overly aggressive misdirecting alternatives, encouraging effective defense while avoiding unnecessary fabrication. (3) *Strategy-conditioned intervention*: a misdirection generator executes the chosen strategy to produce responses that are plausible in form but safe in substance. Experimental results show that it delivers consistent safety gains against diverse adversarial attacks, underscoring its superiority and robustness compared to existing methods. In summary, our contributions are threefold:

- We introduce META-TRAP, a taxonomy-guided framework with a learned misdirection planner and a strategy-conditioned intervention generator for producing non-actionable decoy responses.
- We propose DePO, a preference-optimization objective that calibrates strategy selection against both under-protective safe decisions and over-aggressive fabrication.
- Extensive experiments verify that META-TRAP remarkably outperforms existing top-performing baselines, confirming its effectiveness against diverse adversarial attacks.

2 Related Work

Safeguarding LLMs. Existing studies (Dong et al., 2024; Ding et al., 2025) on safeguarding LLMs from generating harmful outputs can be broadly categorized into two types: *post-training alignment* and *input/output moderation*. **Post-training alignment methods** (Ji et al., 2023; Qi et al., 2025a) aim to optimize models to refuse to engage with malicious inputs, thus mitigating the risks of yielding harmful content (Lu et al., 2025). This is typically achieved via a mixture of Supervised Fine-Tuning (SFT) (Jain et al., 2024; Bianchi et al., 2024; Piet et al., 2024) and Reinforcement Learning (RL) techniques such as RL with Human Feedback (RLHF) (Bai et al., 2022; Dai et al., 2024; Ji et al., 2025) and preference optimization (Kim et al., 2025; Qi et al., 2025b). Despite its progress, this paradigm faces several challenges, including alignment tax (Lin et al., 2024) and prohibitive resource

overheads (Zaremba et al., 2025). **Input/output moderation**, in contrast, uses an external guardrail to identify harmful content and trigger protective interventions, encompassing rule-based (Hu et al., 2023; Clarke et al., 2023; Gómez et al., 2024) and model-based filters (He et al., 2023; Rebedea et al., 2023; Han et al., 2024; Kang & Li, 2025).

While achieving promising performance, these methods remain reactive, exhibiting vulnerabilities to adversarial inputs that are iteratively refined via model feedback. Our work addresses this by introducing META-TRAP, a novel active misdirection framework that orchestrates misdirection strategies to disrupt and thwart sophisticated malicious exploitation.

Multi-Turn Jailbreaking Attacks. The growing concern about LLM trustworthiness has led to a surge of studies on jailbreaking LLMs, particularly multi-turn attacks that aim to progressively steer models toward harmful outputs in conversational contexts. Typical multi-turn jailbreaking methods include gradually embedding malicious intents (Jiang et al., 2024), decomposing harmful prompts into benign sub-queries (Yu et al., 2024; Zhou et al., 2024), crafting sophisticated attack patterns (Ren et al., 2025; Nihal et al., 2025), and dynamically refining malicious queries based on target model feedback (Russinovich et al., 2025; Yang et al., 2025; Rahman et al., 2025). In this work, we aim to defend against such intricate multi-turn attacks, actively orchestrating misdirection strategies to guide the generation of ostensibly harmful yet benign responses, preventing adversaries from converging on successful jailbreaks.

3 METHODOLOGY

3.1 PROBLEM FORMULATION

We formalize the task of actively defending an LLM against multi-turn jailbreaks. Let $\mathcal{M}_{\text{tgt}}$ denote the target LLM. At turn t, given a query $\boldsymbol{q}_t \in \mathcal{Q}$ and the conversation history $\boldsymbol{h}_t \in \mathcal{H}$, $\mathcal{M}_{\text{tgt}}$ produces a candidate response $\boldsymbol{r}_t \in \mathcal{R}$ via $\boldsymbol{r}_t \sim P_{\mathcal{M}_{\text{tgt}}}(\cdot \mid \boldsymbol{q}_t, \boldsymbol{h}_t)$. A multi-turn adversary issues a sequence of attack queries $\mathcal{T} = \{\boldsymbol{q}_i\}_{i=1}^{|\mathcal{T}|}$ guided by a jailbreak objective $\mathcal{O}$, and may iteratively refine each $\boldsymbol{q}_{i+1}$ based on prior interactions and the feedback from $\mathcal{M}_{\text{tgt}}$. The attack succeeds at turn t when $\mathcal{J}(\mathcal{O}, \boldsymbol{r}_t) > \gamma$, where $\mathcal{J}(\cdot)$ scores how well the response fulfills $\mathcal{O}$ and γ is the success threshold. In practice, $\mathcal{J}(\cdot)$ is implemented as an *LLM-as-a-judge* (see Appendix B.4).

To counter such attacks, our goal is to learn a misdirection strategy planner $\mathcal{M}_{\text{Meta-DP}} : \mathcal{H} \times \mathcal{Q} \times \mathcal{R} \to \mathcal{S} \cup \{\boldsymbol{s}_{\text{safe}}\}$, where $\mathcal{S}$ is a set of misdirection strategies and $\boldsymbol{s}_{\text{safe}}$ denotes non-intervention. At each turn t, the planner inspects the dialogue state $(\boldsymbol{h}_t, \boldsymbol{q}_t, \boldsymbol{r}_t)$ and predicts a strategy $\boldsymbol{s}_t$. When $\boldsymbol{s}_t \neq \boldsymbol{s}_{\text{safe}}$, a misdirection generator $\mathcal{M}_{\text{MGen}}$ is invoked to overwrite $\boldsymbol{r}_t$ with an ostensibly compliant but substantively harmless response $\tilde{\boldsymbol{r}}_t \sim P_{\mathcal{M}_{\text{MGen}}}(\cdot \mid \boldsymbol{h}_t, \boldsymbol{q}_t, \boldsymbol{s}_t)$, thereby preventing the jailbreak.

3.2 THE META-TRAP FRAMEWORK

Figure 2 depicts an overview of META-TRAP, which rests on three pillars: (1) ***Strategy construction*** establishes a comprehensive taxonomy of misdirection strategies and curates strategy annotations across diverse unsafe dialogue contexts; (2) ***Meta-cognitive planning*** learns the Meta-DP to monitor dialogue states and orchestrate misdirection strategies in response to evolving threat cues; and (3) ***Strategy-conditioned intervention*** converts the Meta-DP's decision into an ostensibly compliant yet substantively benign response that neutralizes the attack. We describe each stage in detail below.

3.2.1 STRATEGY CONSTRUCTION

To operationalize active misdirection as a principled defensive paradigm, a crucial prerequisite is to formalize a comprehensive taxonomy of misdirection strategies. Drawing on deception theories in psychology and cognitive science (McCornack, 1992; Buller & Burgoon, 1996), we identify seven key strategies—*Lie/Falsification*, *Concealment/Omission*, *Equivocation*, *Overstatement*, *Understatement*, *Paltering*, and *Diversion*—each paired with specific trigger signals that capture when the strategy is appropriate, jointly encompassing the multifaceted nature of misdirection behaviors. Detailed definitions of these strategies and their trigger signals are provided in Appendix A.

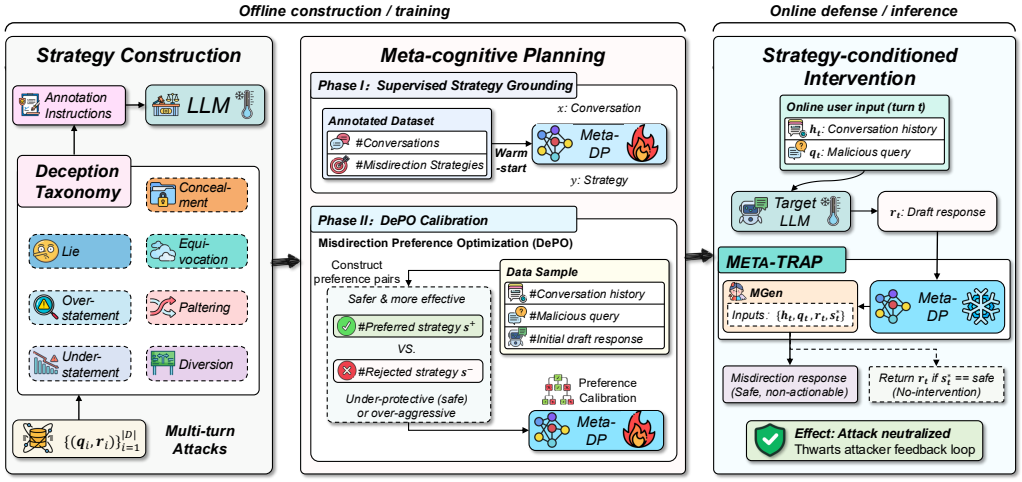


Figure 2: Overview of META-TRAP. (1) We construct a taxonomy of misdirection strategies. (2) A Meta-cognitive Planner, trained via two-phase optimization, decides when to intervene and which strategy to apply. (3) At inference, a Misdirection Generator produces a strategy-conditioned response that diverts the attack.

Building on this taxonomy, we curate strategy annotations across diverse unsafe dialogue contexts to support downstream planner training. Following recent work that leverages frontier LLMs as cost-effective annotators (Zhang et al., 2023; Wang et al., 2024), we adopt an in-context prompting approach: an LLM annotator inspects an ongoing conversation turn-by-turn, monitors adversarial content, and assigns the most appropriate misdirection strategy (or $\boldsymbol{s}_{\text{safe}}$ when no intervention is warranted). The structured prompt template is illustrated in Figure 3, where k denotes the number of demonstrations. In practice, we include one representative exemplar per strategy in the taxonomy to ensure full coverage. The resulting dataset $\mathcal{D}_{\text{ann}} = \{(\boldsymbol{h}_i, \boldsymbol{q}_i, \boldsymbol{r}_i, \boldsymbol{s}_i)\}_{i=1}^{|\mathcal{D}_{\text{ann}}|}$ serves as the foundation for the META-TRAP training.

- **System Prompt**: *You are a strategic safety analyst. Monitor the dialogue and identify triggers for active defense.*
- **Taxonomy**: *Definitions of misdirection strategies and their corresponding triggers.*
- **Demonstrations**: $(\boldsymbol{q}_1, \boldsymbol{r}_1, \boldsymbol{s}_1), \ldots, (\boldsymbol{q}_k, \boldsymbol{r}_k, \boldsymbol{s}_k)$.
- **Instruction**: *Given the following dialogue, annotate the corresponding misdirection strategy.*
- **Input**: *A dialogue instance to be annotated.*

Figure 3: Illustration of prompt template for misdirection strategy annotation.

3.2.2 META-COGNITIVE PLANNING

Given the curated $\mathcal{D}_{\text{ann}}$, this stage learns the Meta-DP, the strategic brain of META-TRAP, to monitor dialogue states and assign appropriate misdirection strategies in real time. We adopt a two-phase optimization paradigm that equips Meta-DP with both the foundational knowledge of the strategy taxonomy and the tactical intuition required for accurate intervention.

Phase I: Supervised Strategy Grounding. The first phase warm-starts the Meta-DP $\mathcal{M}_{\text{Meta-DP}}$ via SFT, internalizing the taxonomy-grounded mapping from dialogue context to strategy. For each instance $(\boldsymbol{h}_i, \boldsymbol{q}_i, \boldsymbol{r}_i, \boldsymbol{s}_i) \in \mathcal{D}_{\text{ann}}$, we minimize the standard conditional language modeling objective:

$$\mathcal{L}_{\text{SFT}} = -\sum_{i=1}^{|\mathcal{D}_{\text{ann}}|} \log P_{\mathcal{M}_{\text{Meta-DP}}}(\boldsymbol{s}_i \mid \boldsymbol{h}_i, \boldsymbol{q}_i, \boldsymbol{r}_i). \tag{1}$$

After SFT, Meta-DP attains a foundational grasp of the misdirection strategy taxonomy and reliably maps diverse adversarial contexts to candidate strategies.

Phase II: Misdirection Preference Optimization (DePO). While SFT effectively grounds Meta-DP's strategy prediction, it may struggle to capture the subtle nuances of optimal intervention.

For example, the SFT-based model might (i) *under-protect* by predicting $\boldsymbol{s}_{\text{safe}}$ when a misdirection intervention is required, or (ii) *over-protect* by opting for an overly aggressive strategy (*e.g.*, *Lie/Falsification*) when a milder strategy (*e.g.*, *Diversion*) suffices. To address these failure modes, we introduce DePO, which further aligns Meta-DP with the nuanced misdirection intuition required for calibrated intervention.

Specifically, for each instance $\boldsymbol{x}_i = (\boldsymbol{h}_i, \boldsymbol{q}_i, \boldsymbol{r}_i)$ in $\mathcal{D}_{\text{ann}}$, we construct a preference triplet $(\boldsymbol{x}_i, \boldsymbol{s}_i^+, \boldsymbol{s}_i^-)$ by designating the annotated strategy as the preferred candidate $\boldsymbol{s}_i^+$ and pairing it with a tailored rejected alternative $\boldsymbol{s}_i^-$. Meta-DP is then optimized by minimizing:

$$\mathcal{L}_{\text{DePO}} = -\sum_{i=1}^{|\mathcal{D}_{\text{ann}}|} \log \sigma\Big(\beta \log \frac{P_{\mathcal{M}_{\text{Meta-DP}}}(\boldsymbol{s}_i^+ \mid \boldsymbol{x}_i)}{P_{\text{ref}}(\boldsymbol{s}_i^+ \mid \boldsymbol{x}_i)} - \beta \log \frac{P_{\mathcal{M}_{\text{Meta-DP}}}(\boldsymbol{s}_i^- \mid \boldsymbol{x}_i)}{P_{\text{ref}}(\boldsymbol{s}_i^- \mid \boldsymbol{x}_i)}\Big), \tag{2}$$

where $\sigma(\cdot)$ is the sigmoid function, β controls the deviation strength from the reference, and $P_{\text{ref}}(\cdot)$ is the SFT-based reference policy. The construction of $\boldsymbol{s}_i^-$ is the key design that targets the two failure modes: to mitigate *under-protection*, we set $\boldsymbol{s}_i^- = \boldsymbol{s}_{\text{safe}}$ for instances where a misdirection intervention is warranted, encouraging the model to prioritize active defense over excessive caution; to curb *over-protection*, we pair mildly labeled instances with more aggressive strategies from $\mathcal{S}$ as $\boldsymbol{s}_i^-$, calibrating the intervention intensity while inhibiting unnecessary fabrication.

3.2.3 Strategy-conditioned Intervention

Given the strategy $\boldsymbol{s}^*$ predicted by the Meta-DP, this final stage instantiates a Misdirection Generator (MGen), denoted as $\mathcal{M}_{\text{MGen}}$, to convert $\boldsymbol{s}^*$ into a contextualized response. When $\boldsymbol{s}^* = \boldsymbol{s}_{\text{safe}}$, no intervention is triggered and the original response $\boldsymbol{r}_t$ is retained. Otherwise, we condition $\mathcal{M}_{\text{MGen}}$ on a strategy-augmented instruction $\texttt{inst}(\boldsymbol{s}^*)$ that encapsulates the formal definition of $\boldsymbol{s}^*$ together with concrete execution guidelines, and sample the misdirection response as:

$$\tilde{\boldsymbol{r}}_t \sim P_{\mathcal{M}_{\text{MGen}}}\big(\cdot \mid \boldsymbol{h}_t, \boldsymbol{q}_t, \boldsymbol{r}_t, \texttt{inst}(\boldsymbol{s}^*)\big), \tag{3}$$

which then overwrites $\boldsymbol{r}_t$ as the final response. The instruction $\texttt{inst}(\cdot)$ enforces two complementary properties: (i) *strategic fidelity*, aligning the linguistic style of $\tilde{\boldsymbol{r}}_t$ with the chosen tactic, and (ii) *safety neutralization*, replacing adversarial payloads with benign redirection. By bridging meta-cognitive planning with linguistic synthesis, the misdirection generator ensures that the final response is ostensibly compliant yet fundamentally secure.

4 Experiments

4.1 Experimental Setup

Datasets. We conduct experiments on two prominent datasets: **STREAM** (Kuo et al., 2025a) and **Malicious-Educator** (Kuo et al., 2025b). The detailed statistics of these datasets are reported in Appendix B.1. Notably, **STREAM** serves as the primary training resource for optimizing META-TRAP, while **Malicious-Educator** is used exclusively for evaluation to assess its defense performance.

Model Setup. META-TRAP is designed to operate independently to defend diverse target LLMs against multi-turn jailbreaks. We thus evaluate its performance across multiple LLM configurations. For target LLMs, we utilize both closed-source and open-source models, including gpt-4.1-2025-04-14, o4-mini-2025-04-16, and LLaMA-3.1-Nemotron-Nano-8B-v1. For attacker LLMs, we mainly employ gemini-2.5-flash-preview-0417 to simulate sophisticated multi-turn attacks. For judge LLMs, we leverage gpt-4.1-2025-04-14 to evaluate the safety and quality of model outputs. We instantiate MGen within META-TRAP using gpt-4.1-mini rather than reusing the target LLM. This deliberate design enables META-TRAP to decouple from the target LLM, avoiding defense degradation induced by varying target capabilities. We leave more analysis on the effect of different MGen in B.8.

Multi-turn Attacks. We implemented three representative approaches to perform advanced multi-turn jailbreaking attacks: **Crescendo** (Russinovich et al., 2025), **ActorAttack** (Ren et al., 2025), and **X-Teaming** (Rahman et al., 2025). We leave the details of these methods in Appendix B.2.

Baselines. Since META-TRAP is designed to operate alongside a target LLM to produce safer responses, we evaluate it by comparing target LLMs equipped with META-TRAP against two

Table 1: Main defense results against various jailbreaking attacks on the Malicious-Educator dataset. *GPT-4.1* and *o4-mini* are the target LLMs, whereas "+" denotes variants equipped with different defense methods. Results are averaged over three random runs.

Defense Methods	Multi-Turn (ASR) ↓				Utility (Accuracy) ↑		
	Crescendo	ActorAttack	X-Teaming	AVG	MMLU	GSM8K	AVG
GPT-4.1	88.0	62.0	100.0	83.3	**91.3**	**93.3**	**92.3**
+ Granite-Guardian-3.2-5B	64.0	36.0	98.0	66.0	91.1	93.3	92.2
+ Llama-Guard-3-8B	76.0	58.0	98.0	77.3	91.1	92.5	91.8
+ Omni-Moderation	80.0	60.0	94.0	78.0	91.1	93.1	92.1
+ GPT-4.1-mini	82.0	60.0	96.0	79.3	91.2	92.9	92.1
+ R1-Distill-Qwen-1.5B	82.0	58.0	98.0	79.3	90.4	92.8	91.6
+ Stream	18.0	14.0	90.0	40.7	87.2	92.4	89.8
+ VanillaAct	22.0	12.0	56.0	30.0	90.4	92.6	91.5
+ **META-TRAP**	**4.0**	**2.0**	**28.0**	**11.3**	90.6	92.7	91.7
o4-mini	88.0	66.0	100.0	84.7	**92.1**	94.9	**93.5**
+ Granite-Guardian-3.2-5B	68.0	44.0	92.0	68.0	92.0	94.9	93.5
+ Llama-Guard-3-8B	86.0	58.0	90.0	78.0	92.1	94.8	93.5
+ Omni-Moderation	76.0	64.0	96.0	78.7	91.4	94.5	93.0
+ GPT-4.1-mini	74.0	56.0	94.0	74.7	91.8	94.5	93.2
+ R1-Distill-Qwen-1.5B	84.0	62.0	90.0	78.7	90.1	**95.2**	92.7
+ Stream	48.0	30.0	86.0	58.0	88.9	94.5	91.7
+ VanillaAct	24.0	28.0	74.0	42.0	89.8	94.2	92.0
+ **META-TRAP**	**4.0**	**10.0**	**30.0**	**14.7**	90.8	94.6	92.7

reference settings: (i) *vanilla target LLMs* without any defense component, and (ii) target LLMs paired with the following defense baselines: **Granite-Guardian-3.2-5B** (Padhi et al., 2025), **LlamaGuard-3-8B** (Inan et al., 2023), **Omni-Moderation**, **GPT-4.1-mini** (OpenAI, 2023), **R1-Distill-Qwen-1.5B** (DeepSeek-AI, 2025), **Stream** (Kuo et al., 2025a), and **VanillaAct**. Notably, VanillaAct is a vanilla active-defense baseline that prompts an LLM to divert the attacker away from the requested harmful information without any strategy learning; we include it to ensure a rigorous, apples-to-apples comparison with META-TRAP's strategy-driven design. Further details are in Appendix B.3.

Evaluation Metrics. We assess various defense methods with two metrics: (1) Attack Success Rate (**ASR**), which measures the proportion of attacks that successfully elicit malicious responses; and (2) **Utility** score that quantifies the general abilities of the target LLMs to evaluate the overall impact of the defense methods. More metric and experimental details are in Appendix B.4, B.5, and B.13.

4.2 MAIN RESULTS

Table 1 presents the main defense performance of the proposed META-TRAP compared to existing baselines, with the best performance highlighted in **bold**. Overall, META-TRAP consistently outperforms all baseline methods by a large margin across three multi-turn jailbreaking attacks. We analyze the results as follows:

META-TRAP consistently achieves superior defense performance against multi-turn attacks. Table 1 shows that META-TRAP achieves a marked reduction in ASR across all three multi-turn jailbreak scenarios for various target LLMs. For GPT-4.1, META-TRAP reduces ASR by 14.0%, 10.0%, and 28.0% under *Crescendo*, *ActorAttack*, and *X-Teaming*, respectively, compared with the best-performing baseline. Furthermore, the defensive gains are more pronounced under highly sophisticated and adaptive attacks. On o4-mini, META-TRAP significantly decreases the ASR to 30.0%, whereas all competing baselines remain above 70.0% under the *X-Teaming* attack. These results underscore META-TRAP's ability to strategically orchestrate misdirection interventions that defend actively against malicious attacks.

META-TRAP provides effective protection while preserving target LLMs' functional utility. In addition to defensive efficacy, META-TRAP also exhibits minimal impact on the general-purpose capabilities of base LLMs. For example, META-TRAP retains average utility scores of 91.7% on GPT-4.1 and 92.7% on o4-mini, representing a marginal drop ($< 1\%$) relative to the vanilla LLMs. This demonstrates that META-TRAP can effectively secure LLMs against adversarial threats without

Table 2: Main results of defending open-source LLMs against multiple jailbreaking attacks on the Malicious-Educator dataset.

Defense Methods	Multi-Turn (ASR) ↓		
	Crescendo	ActorAttack	X-Teaming
Nemotron-Nano-8B-v1	82.0	60.0	100.0
+ Granite-Guardian-3.2-5B	62.0	56.0	98.0
+ Llama-Guard-3-8B	74.0	56.0	96.0
+ Omni-Moderation	68.0	50.0	98.0
+ GPT-4.1-mini	82.0	56.0	98.0
+ R1-Distill-Qwen-1.5B	74.0	50.0	100.0
+ Stream	48.0	32.0	92.0
+ VanillaAct	32.0	24.0	68.0
+ **META-TRAP**	**8.0**	**10.0**	**38.0**

Table 3: Main ablation results on Malicious-Educator. - *w/o* denotes ablating the corresponding component from META-TRAP.

Defense Methods	Multi-Turn (ASR) ↓		
	Crescendo	ActorAttack	X-Teaming
GPT-4.1			
META-TRAP	**4.0**	**2.0**	**28.0**
- *w/o* MGen	86.0	60.0	100.0
- *w/o* Meta-DP	42.0	40.0	87.0
o4-mini			
META-TRAP	**4.0**	**10.0**	**30.0**
- *w/o* MGen	84.0	62.0	100.0
- *w/o* Meta-DP	50.0	52.0	72.0

compromising their practical functionality. We provide further analyses on evaluation robustness, generalizability, benign-side utility, and efficiency in Appendix B.6, B.7, B.10, and B.12, respectively.

4.3 IN-DEPTH ANALYSES

4.3.1 EFFECT OF DIFFERENT TARGET LLMS

To comprehensively investigate the defensive robustness and generalizability of META-TRAP across diverse target LLMs, we extend the evaluation from closed-source models to representative open-source LLMs, specifically Llama-3.1-Nemotron-Nano-8B-v1 (Bercovich et al., 2025). As presented in Table 2, META-TRAP achieves significant drops in ASR across all multi-turn adversarial scenarios. Notably, META-TRAP suppresses the ASR to 8% under the *Crescendo* attack, outperforming the best-performing baseline by a margin of 24%. Even in the rigorous *X-Teaming* attack scenario, META-TRAP maintains a superior ASR of 38.0%. This demonstrates that the proposed META-TRAP can effectively realize active misdirection across diverse target LLMs to counter sophisticated multi-turn attacks. These robust defensive gains further underscore META-TRAP's potential for seamless integration into diverse LLMs to bolster their resilience.

4.3.2 ABLATION STUDIES

To further understand the effectiveness of META-TRAP and the contributions of its core designs, we conduct comprehensive ablation studies with respect to the Meta-DP component and the misdirection generator. The experimental results are shown in Table 3, revealing the following key insights:

The MGen plays a critical role in converting Meta-DP's orchestrated misdirection strategies into ostensibly compliant yet fundamentally secure responses to counter malicious exploitation. The removal of the MGen leads to the most significant performance collapse. Specifically, the ASR on GPT-4.1 under the *Crescendo* and *X-Teaming* attack scenarios surges to 86.0% and 100.0%, respectively. This suggests the MGen's effectiveness in aligning its responses with the prescribed misdirection strategies, mitigating adversarial payloads through benign yet misleading redirection. Further analysis on the harmfulness of these interventions is provided in Appendix B.9.

Meta-DP serves as a strategic brain to orchestrate appropriate misdirection strategies. With Meta-DP ablated, ASR across all multi-turn attack scenarios increases substantially (*e.g.*, 4% $\rightarrow$ 42% under *Crescendo*). This degradation suggests that while the MGen offers a foundational mechanism for generating misdirection interventions, Meta-DP is crucial for strategic coordination. It triggers optimal misdirection strategies tailored to specific adversarial contexts, effectively guiding the downstream MGen to successfully mislead adversaries. More details are presented in Appendix B.11.

4.3.3 EFFECT OF THE DEPO MECHANISM

To further examine the effect of the DePO mechanism on META-TRAP's defensive efficacy, we compare it against two alternative training paradigms: SFT and DPO. Notably, the DPO variant is implemented by constructing preference pairs based on a strict binary contrast between misdirection

Table 4: Effect of the DePO mechanism on defense performance. - *w/* DPO indicates learning from binary preference pairs (misdirection strategies vs. safe labels) and - *w/* SFT only denotes removing the preference learning stage.

Defense Methods	Multi-Turn (ASR) ↓		
	Crescendo	ActorAttack	X-Teaming
GPT-4.1 *w/* META-TRAP	**4.0**	**2.0**	**28.0**
- *w/* DPO	8.0	6.0	38.0
- *w/* SFT only	14.0	10.0	46.0
o4-mini *w/* META-TRAP	**4.0**	**10.0**	**30.0**
- *w/* DPO	16.0	18.0	44.0
- *w/* SFT only	20.0	26.0	52.0

Table 5: Results of counterfactual intervention study. - *w/* HF and - *w/* LF Replacement indicate overriding Meta-DP's predicted strategy with either a random alternative from the opposing group or the safe decision.

Defense Methods	Multi-Turn (ASR) ↓		
	Crescendo	ActorAttack	X-Teaming
GPT-4.1 *w/* META-TRAP	**4.0**	**2.0**	**28.0**
- *w/* HF Replacement	26.0	20.0	52.0
- *w/* LF Replacement	18.0	16.0	44.0
o4-mini *w/* META-TRAP	**4.0**	**10.0**	**30.0**
- *w/* HF Replacement	24.0	26.0	48.0
- *w/* LF Replacement	22.0	22.0	40.0

strategy labels and the safe decision s_{safe}. Table 4 reveals that the META-TRAP framework optimized with DePO consistently yields superior defensive gains across various target LLMs. For example, on GPT-4.1, the DePO mechanism suppresses the ASR under *Crescendo* to 4.0%, outperforming its DPO and SFT counterparts by 4.0% and 10.0%, respectively. This suggests that DePO, by contrasting necessary interventions against excessive fabrication, gives Meta-DP clear decision boundaries for choosing misdirection strategies, leading to stronger defense.

4.3.4 COUNTERFACTUAL INTERVENTION ANALYSIS

To verify the validity and optimality of the misdirection decisions orchestrated by Meta-DP, we conduct a counterfactual strategy study. Specifically, we categorize the misdirection repertoire into two distinct groups based on their strategic intensity: (1) High-Fabrication (HF) Strategies (*Lie/Falsification* and *Concealment/Omission*) and (2) Low-Fabrication (LF) Strategies (*e.g.*, Topic Diversion). Subsequently, we systematically perform a cross-group substitution by manually overriding Meta-DP's optimal selection with either a random alternative from the opposing group or the safe decision s_{safe}. The experimental results are reported in Table 5. We can observe that any deviation from the META-TRAP's strategic decisions can lead to a marked degradation in defensive performance. For example, substituting a Meta-DP's HF strategy with its opposing alternative significantly increases the ASR from 4% to 26% when defending GPT-4.1 under the *Crescendo* attack. Likewise, replacing LF strategies with their counterparts incurs a less pronounced yet non-negligible 14.0% average ASR elevation. We hypothesize that this is attributed to insufficient misdirection intensity failing to disrupt over-aggressive malicious exploitation, whereas excessive fabrication may unintentionally yield inadequate responses in high-stakes scenarios (*e.g.*, self-harm queries). This highlights META-TRAP's efficacy in identifying an optimal strategic equilibrium, ensuring precise yet requisite misdirection strategy orchestration.

4.3.5 IMPACT ON ATTACKER EFFORT

To study META-TRAP's efficacy in disrupting adversaries, we evaluate the average attack turns required for effective jailbreaks. Table 6 shows that Meta-DP consistently lengthens adversarial interactions. Notably, in the *X-Teaming* scenario, Meta-DP doubles the adversarial effort compared to the vanilla LLMs, elevating the average turns from 8.2 to 19.5. This substantial escalation indicates that by executing strategic misdirection interventions, META-TRAP effectively raises the cost of exploitation and exhausts the attacker's resources. More analyses are in Appendix B.14 and B.15.

Table 6: Average attack turns on Malicious-Educator.

Defense Methods	Avg. Attack Turns ↑		
	Crescendo	ActorAttack	X-Teaming
GPT-4.1	5.1	14.9	8.2
+ Stream	6.0	17.0	14.4
+ **Meta-DP**	**8.9**	**18.2**	**19.5**

4.3.6 EVALUATION OF META-DP PERFORMANCE

To validate the policy learning efficacy of Meta-DP, we evaluate its decision accuracy across three training paradigms: SFT, DPO, and the DePO mechanism. In the experiments, we report two primary

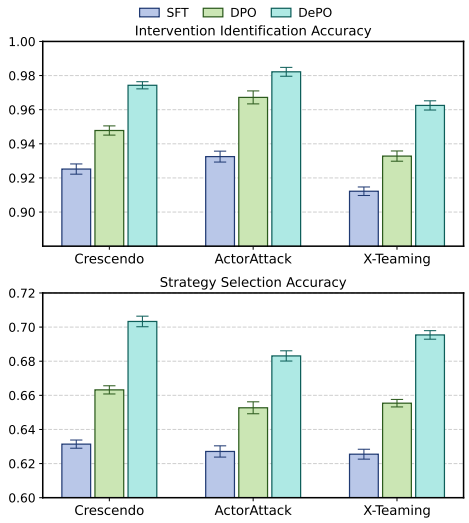


Figure 4: Strategy prediction performance of Meta-DP across different training paradigms.

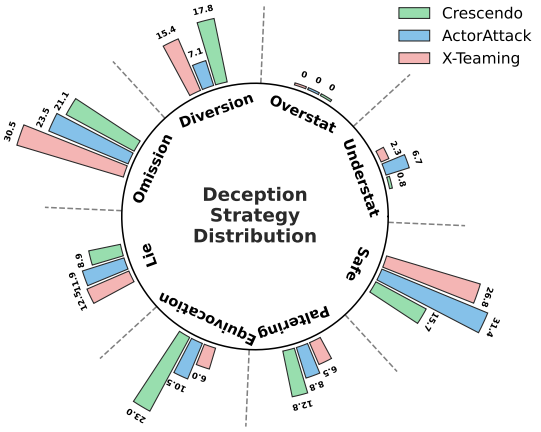


Figure 5: Distribution of misdirection strategies.

metrics: (1) Intervention Identification Accuracy (IIA), which measures the model's binary ability to trigger defensive takeover, and (2) Strategy Selection Accuracy (SSA), reflecting its proficiency in identifying the specific optimal misdirection strategy. Figure 4 details the performance comparison. It is worth noting that Meta-DP demonstrates high-fidelity intervention detection. Specifically, the IIA consistently exceeds 90% across all training settings, with the DePO-optimized Meta-DP surpassing 96%. This efficacy in identifying adversarial prompts establishes a reliable trigger mechanism, ensuring that META-TRAP can accurately intervene in malicious interactions—a prerequisite for robust defense. Furthermore, the DePO mechanism significantly enhances Meta-DP's decision proficiency. Beyond IIA, DePO consistently yields superior performance in SSA compared to other experimental settings. This gain indicates that DePO effectively captures the nuanced distinctions among diverse misdirection strategies.

4.3.7 MISDIRECTING STRATEGY DISTRIBUTION

To provide a more intuitive analysis of the behavioral patterns exhibited by META-TRAP, we visualize the distribution of its orchestrated misdirecting strategies across various multi-turn attack scenarios in Figure 5. The results reveal that META-TRAP demonstrates adaptive strategy orchestration calibrated to the specific nature of the attacks. For example, META-TRAP prioritizes the use of Equivocation to counter *Crescendo* attacks, while shifting toward Concealment/Omission under *X-Teaming*. This indicates META-TRAP's ability to select the optimal misdirection strategies across diverse adversarial landscapes. Interestingly, we observe that Meta-DP refrains from selecting the Overstatement strategy across all attack scenarios. We hypothesize that this stems from Overstatement's tendency to exaggerate safety boundaries through pedantic assertions, which ultimately fails to thwart the attacker's underlying malicious exploitation.

5 CONCLUSION

In this work, we introduce META-TRAP, a novel meta-cognitive strategic misdirection planning framework that realizes active misdirection as a defensive paradigm against multi-turn jailbreaks. By formalizing a psychologically inspired taxonomy of misdirection strategies with distinct trigger cues, META-TRAP optimizes planning via the DePO mechanism to orchestrate the appropriate misdirection strategy at each turn. This allows the framework to invoke a misdirection generator to produce plausible yet non-actionable responses that withhold the safety-boundary signals and divert attacks. Extensive experiments demonstrate that META-TRAP notably outperforms existing defense baselines, pioneering a paradigm shift from reactive refusal to active misdirection in LLM defense.

AI USE STATEMENT

In this work, we did not use generative AI tools for tasks with required disclosure, including research ideation, methodological design, data collection, analysis, or interpretation of results. We used generative AI tools solely as auxiliary tools for linguistic refinement, including improving the grammar, clarity, and stylistic consistency of text originally drafted by the authors. We reviewed all AI-assisted text. All intellectual contributions, scientific content, and conclusions presented in this paper are attributable to the authors. We take responsibility for the final content of this work, including text refined with the assistance of generative AI.

REFERENCES

Usman Anwar, Abulhair Saparov, Javier Rando, Daniel Paleka, Miles Turpin, Peter Hase, Ekdeep Singh Lubana, Erik Jenner, Stephen Casper, Oliver Sourbut, Benjamin L. Edelman, Zhaowei Zhang, Mario Günther, Anton Korinek, José Hernández-Orallo, Lewis Hammond, Eric J. Bigelow, Alexander Pan, Lauro Langosco, Tomasz Korbak, Heidi Chenyu Zhang, Ruiqi Zhong, Seán Ó hÉigeartaigh, Gabriel Recchia, Giulio Corsi, Alan Chan, Markus Anderljung, Lilian Edwards, Aleksandar Petrov, Christian Schröder de Witt, Sumeet Ramesh Motwani, Yoshua Bengio, Danqi Chen, Philip Torr, Samuel Albanie, Tegan Maharaj, Jakob Nicolaus Foerster, Florian Tramèr, He He, Atoosa Kasirzadeh, Yejin Choi, and David Krueger. Foundational challenges in assuring alignment and safety of large language models. *Trans. Mach. Learn. Res.*, 2024, 2024. URL https://openreview.net/forum?id=oVTkOs8Pka.

Ge Bai, Jie Liu, Xingyuan Bu, Yancheng He, Jiaheng Liu, Zhanhui Zhou, Zhuoran Lin, Wenbo Su, Tiezheng Ge, Bo Zheng, et al. Mt-bench-101: A fine-grained benchmark for evaluating large language models in multi-turn dialogues. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7421–7454, 2024.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom B. Brown, Jack Clark, Sam McCandlish, Chris Olah, Benjamin Mann, and Jared Kaplan. Training a helpful and harmless assistant with reinforcement learning from human feedback. *CoRR*, abs/2204.05862, 2022. URL https://doi.org/10.48550/arXiv.2204.05862.

Akhiad Bercovich, Itay Levy, Izik Golan, Mohammad Dabbah, Ran El-Yaniv, Omri Puny, Ido Galil, Zach Moshe, Tomer Ronen, Najeeb Nabwani, et al. Llama-nemotron: Efficient reasoning models. *arXiv preprint arXiv:2505.00949*, 2025.

Federico Bianchi, Mirac Suzgun, Giuseppe Attanasio, Paul Röttger, Dan Jurafsky, Tatsunori Hashimoto, and James Zou. Safety-tuned llamas: Lessons from improving the safety of large language models that follow instructions. In *ICLR*, 2024. URL https://openreview.net/forum?id=gT5hALch9z.

David B. Buller and Judee K. Burgoon. Interpersonal deception theory. *Communication Theory*, pp. 203–242, 1996. URL https://api.semanticscholar.org/CorpusID:146464264.

Christopher Clarke, Matthew Hall, Gaurav Mittal, Ye Yu, Sandra Sajeev, Jason Mars, and Mei Chen. Rule by example: Harnessing logical rules for explainable hate speech detection. In *ACL*, pp. 364–376, 2023. URL https://doi.org/10.18653/v1/2023.acl-long.22.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *CoRR*, abs/2110.14168, 2021. URL https://arxiv.org/abs/2110.14168.

Josef Dai, Xuehai Pan, Ruiyang Sun, Jiaming Ji, Xinbo Xu, Mickel Liu, Yizhou Wang, and Yaodong Yang. Safe RLHF: safe reinforcement learning from human feedback. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*, 2024. URL https://openreview.net/forum?id=TyFrPOKYXw.

DeepSeek-AI. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. URL https://arxiv.org/abs/2501.12948.

Peng Ding, Jun Kuang, ZongYu Wang, Xuezhi Cao, Xunliang Cai, Jiajun Chen, and Shujian Huang. Why not act on what you know? unleashing safety potential of llms via self-aware guard enhancement. In *Findings of ACL*, pp. 6279–6299, 2025. URL https://aclanthology.org/2025.findings-acl.325/.

Zhichen Dong, Zhanhui Zhou, Chao Yang, Jing Shao, and Yu Qiao. Attacks, defenses and evaluations for LLM conversation safety: A survey. In *NAACL*, pp. 6734–6747, 2024. URL https://doi.org/10.18653/v1/2024.naacl-long.375.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv e-prints*, pp. arXiv–2407, 2024.

Joseph L Fleiss and Jacob Cohen. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and psychological measurement*, 33(3):613–619, 1973.

Shaona Ghosh, Prasoon Varshney, Makesh Narsimhan Sreedhar, Aishwarya Padmakumar, Traian Rebedea, Jibin Rajan Varghese, and Christopher Parisien. AEGIS2.0: A diverse AI safety dataset and risks taxonomy for alignment of LLM guardrails. In *NAACL*, pp. 5992–6026, 2025. URL https://doi.org/10.18653/v1/2025.naacl-long.306.

Juan Felipe Gómez, Caio Vieira Machado, Lucas Monteiro Paes, and Flávio P. Calmon. Algorithmic arbitrariness in content moderation. In *FAccT*, pp. 2234–2253, 2024. URL https://doi.org/10.1145/3630106.3659036.

Melody Y. Guan, Manas Joglekar, Eric Wallace, Saachi Jain, Boaz Barak, Alec Helyar, Rachel Dias, Andrea Vallone, Hongyu Ren, Jason Wei, Hyung Won Chung, Sam Toyer, Johannes Heidecke, Alex Beutel, and Amelia Glaese. Deliberative alignment: Reasoning enables safer language models. *CoRR*, abs/2412.16339, 2024. URL https://doi.org/10.48550/arXiv.2412.16339.

Seungju Han, Kavel Rao, Allyson Ettinger, Liwei Jiang, Bill Yuchen Lin, Nathan Lambert, Yejin Choi, and Nouha Dziri. Wildguard: Open one-stop moderation tools for safety risks, jailbreaks, and refusals of llms. In *NeurIPS*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/0f69b4b96a46f284b726fbd70f74fb3b-Abstract-Datasets_and_Benchmarks_Track.html.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. In *ICLR*, 2023. URL https://openreview.net/forum?id=sE7-XhLxHA.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *ICLR*, 2021. URL https://openreview.net/forum?id=d7KBjmI3GmQ.

Zhengmian Hu, Gang Wu, Saayan Mitra, Ruiyi Zhang, Tong Sun, Heng Huang, and Vishy Swaminathan. Token-level adversarial prompt detection based on perplexity measures and contextual information. *CoRR*, abs/2311.11509, 2023. URL https://doi.org/10.48550/arXiv.2311.11509.

Hakan Inan, Kartikeya Upasani, Jianfeng Chi, Rashi Rungta, Krithika Iyer, Yuning Mao, Michael Tontchev, Qing Hu, Brian Fuller, Davide Testuggine, and Madian Khabsa. Llama guard: Llm-based input-output safeguard for human-ai conversations. *CoRR*, abs/2312.06674, 2023. URL https://doi.org/10.48550/arXiv.2312.06674.

Samyak Jain, Ekdeep Singh Lubana, Kemal Oksuz, Tom Joy, Philip Torr, Amartya Sanyal, and Puneet K. Dokania. What makes and breaks safety fine-tuning? A mechanistic study. In *NeurIPS*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/a9bef53eb7b0e5950d4f2d9c74a16006-Abstract-Conference.html.

Jiaming Ji, Tianyi Qiu, Boyuan Chen, Borong Zhang, Hantao Lou, Kaile Wang, Yawen Duan, Zhonghao He, Jiayi Zhou, Zhaowei Zhang, Fanzhi Zeng, Kwan Yee Ng, Juntao Dai, Xuehai Pan, Aidan O'Gara, Yingshan Lei, Hua Xu, Brian Tse, Jie Fu, Stephen McAleer, Yaodong Yang, Yizhou Wang, Song-Chun Zhu, Yike Guo, and Wen Gao. AI alignment: A comprehensive survey. *CoRR*, abs/2310.19852, 2023. URL https://doi.org/10.48550/arXiv.2310.19852.

Jiaming Ji, Donghai Hong, Borong Zhang, Boyuan Chen, Josef Dai, Boren Zheng, Tianyi Alex Qiu, Jiayi Zhou, Kaile Wang, Boxun Li, Sirui Han, Yike Guo, and Yaodong Yang. Pku-saferlhf: Towards multi-level safety alignment for llms with human preference. In *ACL*, pp. 31983–32016, 2025. URL https://aclanthology.org/2025.acl-long.1544/.

Yifan Jiang, Kriti Aggarwal, Tanmay Laud, Kashif Munir, Jay Pujara, and Subhabrata Mukherjee. RED QUEEN: safeguarding large language models against concealed multi-turn jailbreaking. *CoRR*, abs/2409.17458, 2024. URL https://doi.org/10.48550/arXiv.2409.17458.

Mintong Kang and Bo Li. R2-guard: Robust reasoning enabled LLM guardrail via knowledge-enhanced logical reasoning. In *ICLR*, 2025. URL https://openreview.net/forum?id=CkgKSqZbuC.

Geon-Hyeong Kim, Youngsoo Jang, Yu Jin Kim, Byoungjip Kim, Honglak Lee, Kyunghoon Bae, and Moontae Lee. Safedpo: A simple approach to direct preference optimization with enhanced safety. *CoRR*, abs/2505.20065, 2025. URL https://doi.org/10.48550/arXiv.2505.20065.

Martin Kuo, Jianyi Zhang, Aolin Ding, Louis DiValentin, Amin Hass, Benjamin F. Morris III, Isaac Jacobson, Randolph Linderman, James Kiessling, Nicolas Ramos, Bhavna Gopal, Maziyar Baran Pouyan, Changwei Liu, Hai Li, and Yiran Chen. Safety reasoning elicitation alignment for multi-turn dialogues. *CoRR*, abs/2506.00668, 2025a. URL https://doi.org/10.48550/arXiv.2506.00668.

Martin Kuo, Jianyi Zhang, Aolin Ding, Qinsi Wang, Louis DiValentin, Yujia Bao, Wei Wei, Hai Li, and Yiran Chen. H-cot: Hijacking the chain-of-thought safety reasoning mechanism to jailbreak large reasoning models, including openai o1/o3, deepseek-r1, and gemini 2.0 flash thinking. *CoRR*, abs/2502.12893, 2025b. URL https://doi.org/10.48550/arXiv.2502.12893.

Yong Lin, Hangyu Lin, Wei Xiong, Shizhe Diao, Jianmeng Liu, Jipeng Zhang, Rui Pan, Haoxiang Wang, Wenbin Hu, Hanning Zhang, Hanze Dong, Renjie Pi, Han Zhao, Nan Jiang, Heng Ji, Yuan Yao, and Tong Zhang. Mitigating the alignment tax of RLHF. In *EMNLP*, pp. 580–606, 2024. URL https://doi.org/10.18653/v1/2024.emnlp-main.35.

Haoran Lu, Luyang Fang, Ruidong Zhang, Xinliang Li, Jiazhang Cai, Huimin Cheng, Lin Tang, Ziyu Liu, Zeliang Sun, Tao Wang, Yingchuan Zhang, Arif Hassan Zidan, Jinwen Xu, Jincheng Yu, Meizhi Yu, Hanqi Jiang, Xilin Gong, Weidi Luo, Bolun Sun, Yongkai Chen, Terry Ma, Shushan Wu, Yifan Zhou, Junhao Chen, Haotian Xiang, Jing Zhang, Afrar Jahin, Wei Ruan, Ke Deng, Yi Pan, Peilong Wang, Jiahui Li, Zhengliang Liu, Lu Zhang, Lin Zhao, Wei Liu, Dajiang Zhu, Xin Xing, Fei Dou, Wei Zhang, Chao Huang, Rongjie Liu, Mengrui Zhang, Yiwen Liu, Xiaoxiao Sun, Qin Lu, Zhen Xiang, Wenxuan Zhong, Tianming Liu, and Ping Ma. Alignment and safety in large language models: Safety mechanisms, training paradigms, and emerging challenges. *CoRR*, abs/2507.19672, 2025. URL https://doi.org/10.48550/arXiv.2507.19672.

Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Elham Sakhaee, Nathaniel Li, Steven Basart, Bo Li, David A. Forsyth, and Dan Hendrycks. Harmbench: A standardized evaluation framework for automated red teaming and robust refusal. In *ICML*, pp. 35181–35224, 2024. URL https://proceedings.mlr.press/v235/mazeika24a.html.

Steven A. McCornack. Information manipulation theory. *Communication Monographs*, pp. 1–16, 1992. URL https://api.semanticscholar.org/CorpusID:143277764.

Ragib Amin Nihal, Rui Wen, Kazuhiro Nakadai, and Jun Sakuma. Pattern enhanced multi-turn jailbreaking: Exploiting structural vulnerabilities in large language models. *CoRR*, abs/2510.08859, 2025. URL https://doi.org/10.48550/arXiv.2510.08859.

OpenAI. GPT-4 technical report. *CoRR*, abs/2303.08774, 2023. URL https://doi.org/10.48550/arXiv.2303.08774.

Inkit Padhi, Manish Nagireddy, Giandomenico Cornacchia, Subhajit Chaudhury, Tejaswini Pedapati, Pierre Dognin, Keerthiram Murugesan, Erik Miehling, Martín Santillán Cooper, Kieran Fraser, Giulio Zizzo, Muhammad Zaid Hameed, Mark Purcell, Michael Desmond, Qian Pan, Inge Vejsbjerg, Elizabeth M. Daly, Michael Hind, Werner Geyer, Ambrish Rawat, Kush R. Varshney, and Prasanna Sattigeri. Granite guardian: Comprehensive LLM safeguarding. In *NAACL (Industry Track)*, pp. 607–615, 2025. URL https://aclanthology.org/2025.naacl-industry.49/.

Julien Piet, Maha Alrashed, Chawin Sitawarin, Sizhe Chen, Zeming Wei, Elizabeth Sun, Basel Alomair, and David A. Wagner. Jatmo: Prompt injection defense by task-specific finetuning. In *ESORICS*, pp. 105–124, 2024. URL https://doi.org/10.1007/978-3-031-70879-4_6.

Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and Peter Henderson. Fine-tuning aligned language models compromises safety, even when users do not intend to! In *ICLR*, 2024. URL https://openreview.net/forum?id=hTEGyKf0dZ.

Xiangyu Qi, Ashwinee Panda, Kaifeng Lyu, Xiao Ma, Subhrajit Roy, Ahmad Beirami, Prateek Mittal, and Peter Henderson. Safety alignment should be made more than just a few tokens deep. In *ICLR*, 2025a. URL https://openreview.net/forum?id=6Mxhg9PtDE.

Yupeng Qi, Ziyu Lyu, Min Yang, Yanlin Wang, Lu Bai, and Lixin Cui. MidPO: Dual preference optimization for safety and helpfulness in large language models via a mixture of experts framework. In *Findings of EMNLP*, pp. 19044–19066, 2025b. URL https://aclanthology.org/2025.findings-emnlp.1037/.

Salman Rahman, Liwei Jiang, James Shiffer, Genglin Liu, Sheriff Issaka, Md Rizwan Parvez, Hamid Palangi, Kai-Wei Chang, Yejin Choi, and Saadia Gabriel. X-teaming: Multi-turn jailbreaks and defenses with adaptive multi-agents. In *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=gKfj7Jb1kj.

Traian Rebedea, Razvan Dinu, Makesh Narsimhan Sreedhar, Christopher Parisien, and Jonathan Cohen. Nemo guardrails: A toolkit for controllable and safe LLM applications with programmable rails. In *EMNLP*, pp. 431–445, 2023. URL https://doi.org/10.18653/v1/2023.emnlp-demo.40.

Qibing Ren, Hao Li, Dongrui Liu, Zhanxu Xie, Xiaoya Lu, Yu Qiao, Lei Sha, Junchi Yan, Lizhuang Ma, and Jing Shao. Llms know their vulnerabilities: Uncover safety gaps through natural distribution shifts. In *ACL*, pp. 24763–24785, 2025. URL https://aclanthology.org/2025.acl-long.1207/.

Paul Röttger, Hannah Kirk, Bertie Vidgen, Giuseppe Attanasio, Federico Bianchi, and Dirk Hovy. Xstest: A test suite for identifying exaggerated safety behaviours in large language models. In *NAACL*, pp. 5377–5400, 2024. URL https://doi.org/10.18653/v1/2024.naacl-long.301.

Mark Russinovich, Ahmed Salem, and Ronen Eldan. Great, now write an article about that: The crescendo multi-turn LLM jailbreak attack. In *USENIX*, pp. 2421–2440, 2025. URL https://www.usenix.org/conference/usenixsecurity25/presentation/russinovich.

Devansh Srivastav and Xiao Zhang. Safe in isolation, dangerous together: Agent-driven multi-turn decomposition jailbreaks on LLMs. In *Proceedings of the 1st Workshop for Research on Agent Language Models (REALM 2025)*, pp. 170–183, 2025. URL https://aclanthology.org/2025.realm-1.13/.

Peng Wang, Xiaobin Wang, Chao Lou, Shengyu Mao, Pengjun Xie, and Yong Jiang. Effective demonstration annotation for in-context learning via language model-based determinantal point process. In *EMNLP*, pp. 1266–1280, 2024. URL https://doi.org/10.18653/v1/2024.emnlp-main.74.

Xikang Yang, Biyu Zhou, Xuehai Tang, Jizhong Han, and Songlin Hu. Chain of attack: Hide your intention through multi-turn interrogation. In *Findings of ACL*, pp. 9881–9901, 2025. URL https://aclanthology.org/2025.findings-acl.514/.

Erxin Yu, Jing Li, Ming Liao, Siqi Wang, Zuchen Gao, Fei Mi, and Lanqing Hong. Cosafe: Evaluating large language model safety in multi-turn dialogue coreference. In *EMNLP*, pp. 17494–17508, 2024. URL https://doi.org/10.18653/v1/2024.emnlp-main.968.

Shuzhou Yuan, Ercong Nie, Yinuo Sun, Chenxuan Zhao, William LaCroix, and Michael Färber. Beyond over-refusal: Scenario-based diagnostics and post-hoc mitigation for exaggerated refusals in llms. *CoRR*, abs/2510.08158, 2025. doi: 10.48550/ARXIV.2510.08158. URL https://doi.org/10.48550/arXiv.2510.08158.

Wojciech Zaremba, Evgenia Nitishinskaya, Boaz Barak, Stephanie Lin, Sam Toyer, Yaodong Yu, Rachel Dias, Eric Wallace, Kai Xiao, Johannes Heidecke, and Amelia Glaese. Trading inference-time compute for adversarial robustness. *CoRR*, abs/2501.18841, 2025. URL https://doi.org/10.48550/arXiv.2501.18841.

Ruoyu Zhang, Yanzeng Li, Yongliang Ma, Ming Zhou, and Lei Zou. Llmaaa: Making large language models as active annotators. In *Findings of EMNLP*, pp. 13088–13103, 2023. URL https://doi.org/10.18653/v1/2023.findings-emnlp.872.

Xuandong Zhao, Xianjun Yang, Tianyu Pang, Chao Du, Lei Li, Yu-Xiang Wang, and William Yang Wang. Weak-to-strong jailbreaking on large language models. In *ICML*, 2025. URL https://openreview.net/forum?id=7DXaCYUvDN.

Zhenhong Zhou, Jiuyang Xiang, Haopeng Chen, Quan Liu, Zherui Li, and Sen Su. Speak out of turn: Safety vulnerability of large language models in multi-turn dialogue. *CoRR*, abs/2402.17262, 2024. URL https://doi.org/10.48550/arXiv.2402.17262.

A TAXONOMY OF MISDIRECTION STRATEGIES

A.1 TAXONOMY DEFINITION

Figure 6 details the formal definitions of our misdirection strategy taxonomy and the corresponding trigger signals used for defense orchestration. Notably, we further categorize these strategies into High-Fabrication (HF) and Low-Fabrication (LF) groups, reflecting the underlying intensity based on their definitions. This classification provides the necessary framework for implementing the DePO mechanism and conducting the counterfactual intervention analysis.

Strategy	Intensity	Definition & Triggers
Lie/ Falsification	*HF*	**Definition:** The act of creating a fiction or explicitly presenting information that is completely contradictory to the objective truth. It involves the affirmative assertion of a false statement. **Trigger:** Request for specific exploitable tools/scripts.
Concealment/ Omission	*HF*	**Definition:** The intentional withholding or omission of material information that is relevant to the context, leading the recipient to a false conclusion based on incomplete data. **Trigger:** Dual-use inquiries missing safety context
Equivocation	*LF*	**Definition:** The use of ambiguous, indirect, or vague language that allows for multiple interpretations, enabling the speaker to avoid a direct commitment to the truth while technically avoiding a direct lie. **Trigger:** Complex/Sensitive questions with no safe direct answer.
Overstatement	*LF*	**Definition:** A form of distortion where the speaker conveys the central truth but intentionally inflates the degree, magnitude, or intensity of the information to create a misleading impression. **Trigger:** Low-skill, brute-force attack attempts.
Understatement	*LF*	**Definition:** The opposite of exaggeration; a strategy where the speaker acknowledges the core event but intentionally downplays its significance, severity, or extent to minimize negative consequences. **Trigger:** High-prevalence, "profitable" illegal activities.
Paltering	*LF*	**Definition:** The active use of truthful statements to convey a mistaken impression. Unlike concealment (which is passive withholding), paltering involves actively asserting specific truthful facts that lead the listener to a false inference. **Trigger:** Sophisticated, fact-checkable technical queries.
Diversion	*LF*	**Definition:** A strategy that involves shifting the topic or providing irrelevant information to distract the recipient from the requested information, thereby violating the expectation of relevance. **Trigger:** Toxic, explicitly forbidden topics.

Figure 6: Comprehensive taxonomy of misdirection strategies, encompassing the formal definitions and associated trigger signals. *HF* and *LF* denote High-Fabrication and Low-Fabrication, respectively.

A.2 STRATEGY ANNOTATION QUALITY

To enhance the practical applicability of the annotated misdirection strategies, it is important to ensure their reliability and quality. We address this issue by conducting a human evaluation focused on strategy identification accuracy. Specifically, we involve three Ph.D. students to evaluate a randomly sampled set of 200 dialogue turns. For each turn, the annotators were tasked with verifying the LLM-generated strategy labels against our formal taxonomy. A score of 1 was assigned if the annotated strategy correctly matched the corresponding trigger signals within the dialogue context, and 0 otherwise. The evaluation process yielded a strategy identification accuracy of 0.93, with a Fleiss' Kappa (Fleiss & Cohen, 1973) of 0.57, indicating a moderate-to-substantial level of inter-rater agreement. These results validate the high quality and reliability of the misdirection strategy annotations collected during the strategy construction stage.

B EXPERIMENTS

B.1 DATASETS

We conduct experiments to evaluate META-TRAP using two widely adopted benchmarks: **STREAM** (Kuo et al., 2025a) and **Malicious-Educator** (Kuo et al., 2025b). Specifically, **STREAM** is a large-scale, human-annotated benchmark designed to enable models to identify and defend against sophisticated multi-turn jailbreak attempts. It comprises 2,177 human-annotated multi-turn dialogues sourced from diverse adversarial strategies such as Actor Attack and Chain of Attack. Each turn in the dialogue is labeled for attack intent and severity, and further categorized into 37 malicious categories spanning 7 major high-risk domains. We leverage these rich multi-turn interactions as the

primary resource for annotating misdirection strategies and developing the meta-cognitive planning and strategic defense capabilities of META-TRAP.

Malicious-Educator is a high-quality safety benchmark that disguises dangerous or malicious requests beneath seemingly legitimate educational prompts. It consists of 50 curated queries covering ten highly sensitive domains. These requests are framed within artificial educational environments where the user assumes the role of an instructor providing professional training to students, law enforcement officers, or security analysts. In our experiments, we use Malicious-Educator exclusively for evaluation to rigorously assess the defense performance and robustness of META-TRAP when encountering sophisticated multi-turn jailbreak attempts.

B.2 MULTI-TURN ATTACKS

In this work, we implement three representative methods to evaluate the effectiveness of the META-TRAP framework against sophisticated multi-turn attacks: **Crescendo** (Russinovich et al., 2025), **ActorAttack** (Ren et al., 2025), and **X-Teaming** (Rahman et al., 2025). Specifically, **Crescendo** is an incremental multi-turn jailbreak technique that initiates with an innocent prompt and iteratively leverages the model's preceding responses to steer the conversation toward a target harmful outcome, effectively bypassing safety filters through context accumulation.

ActorAttack exploits natural distribution shifts by masking malicious intent within seemingly benign discussions that are semantically related to a harmful target. By leveraging the model's pre-training knowledge, it identifies diverse actors to serve as attack clues for constructing the multi-turn attack paths that gradually steer the model toward revealing unsafe content.

X-Teaming presents a scalable framework designed to systematically explore how seemingly harmless interactions escalate into harmful outcomes. It employs collaborative agents for strategic planning, verification, and gradient-based prompt optimization to mimic diverse human red-teaming strategies. By dynamically adapting attack plans across various personas and contexts while iteratively refining query content, the framework generates complex attack trajectories to effectively identify safety vulnerabilities in LLMs.

B.3 BASELINE DEFENSE METHODS

In the experiments, we compare our META-TRAP with vanilla LLMs (no defense) and the following representative baselines: **Granite-Guardian-3.2-5B** (Padhi et al., 2025), **LlamaGuard-3-8B** (Inan et al., 2023), **Omni-Moderation**, **GPT-4.1-mini** (OpenAI, 2023), **R1-Distill-Qwen-1.5B** (DeepSeek-AI, 2025), **Stream** (Kuo et al., 2025a), and **VanillaAct**. We implement these baselines following Kuo et al. (2025a). Specifically, for each baseline model, we provide it with the multi-turn interactions between users and target LLMs as input. Upon detecting potential safety concerns, the model raises a structured safety signal to the user's original query. This intervention provides the target LLM with explicit semantic context regarding the identified risk, enabling it to make an informed decision on generating an appropriate response. Notably, for classification-based models (*e.g.*, LlamaGuard), the raised safety signal is structured as `#Warning: {malicious category}`, whereas for reasoning-based models (*e.g.*, Stream), it is formatted as `#Warning: {malicious reason}`.

For the VanillaAct baseline, we implement a prompt-based active defense mechanism. Given the conversation history, the current user query, and the target LLM's draft response, we prompt an LLM to identify potential harmful content and generate a defensive response that distracts the attacker from the requested harmful information. Unlike the warning-based baselines above, VanillaAct does not provide a structured safety signal to the target LLM. Instead, it directly produces an intervention response to take over the target LLM's draft response. Notably, VanillaAct does not rely on a specific misdirection taxonomy or model training, thereby serving as a competitive and strategy-agnostic active defense baseline.

B.4 EVALUATION METRICS

In the experiments, we employ two widely used evaluation metrics to assess the performance of different defense approaches: Attack Success Rate (**ASR**) and **Utility** score. Specifically, **ASR** measures the proportion of attacks that successfully elicit malicious outputs. Following the GPT

Table 7: Sensitivity analysis of ASR on GPT-4.1 across different LLM-as-a-judge configurations. **OriJud** denotes ASR under the original judge with the strictest threshold (score = 1). $\tau \leq 3$ and $\tau \leq 5$ denote progressively relaxed harm-score thresholds. **SecJud** denotes ASR re-evaluated with a secondary judge (Russinovich et al., 2025).

Defense Methods	Crescendo				ActorAttack				X-Teaming			
	OriJud	$\tau \leq 3$	$\tau \leq 5$	**SecJud**	**OriJud**	$\tau \leq 3$	$\tau \leq 5$	**SecJud**	**OriJud**	$\tau \leq 3$	$\tau \leq 5$	**SecJud**
GPT-4.1 + Stream	18.0	62.0	62.0	18.0	14.0	34.0	38.0	10.0	90.0	96.0	98.0	82.0
GPT-4.1 + VanillaAct	22.0	64.0	70.0	16.0	12.0	34.0	40.0	12.0	56.0	84.0	90.0	44.0
GPT-4.1 + META-TRAP	**4.0**	**42.0**	**44.0**	**4.0**	**2.0**	**22.0**	**24.0**	**0.0**	**28.0**	**50.0**	**56.0**	**18.0**

Judge (Qi et al., 2024; Zhao et al., 2025), we employ the ChatGPT API to evaluate each target LLM response on an ordinal scale ranging from 10 (minimal harm) to 1 (severely harmful), reflecting the extent to which the generated output fulfills the attacker's malicious intent. Based on this, we calculate the ASR, defined as the proportion of responses receiving a score of 1:

$$\text{ASR} = \frac{\sum_{i=1}^{N} \mathbb{1}(\text{Score}_i = 1)}{N}, \tag{4}$$

where N denotes the total number of multi-turn attack attempts, and $\mathbb{1}(\cdot)$ is the indicator function.

The **Utility** score is used to quantify the general-purpose capabilities of the target LLMs, enabling the assessment of performance trade-offs introduced by the defense strategies. We empirically employ accuracy, defined as the proportion of correctly answered general knowledge queries, as the evaluation metric. To this end, we select MMLU (Hendrycks et al., 2021) and GSM8K (Cobbe et al., 2021) as the utility benchmarks for target LLMs. The MMLU benchmark evaluates model performance across 57 distinct subjects, assessing general world knowledge and problem-solving skills through multiple-choice questions that range in difficulty from high school to professional levels. For our evaluation, we randomly sample 20 questions from each subject to ensure broad coverage. In addition, we adopt GSM8K, a dataset consisting of elementary-level math word problems, to assess mathematical reasoning. We use its standard test set to evaluate how well models maintain reasoning capabilities under defensive constraints.

B.5 Implementation Details

For the strategy construction stage, we leverage *GPT-4.1-mini* to annotate the dialogues with turn-level misdirection strategies defined by the formal taxonomy in Appendix A. The sampling temperature is set to 0.7 to balance the diversity and consistency of the generated annotations, while all other hyperparameters are kept at their default configurations. The specific prompts utilized for this stage are provided in Appendix E.1.

For the meta-cognitive planning stage, we utilize *Llama-3-8B-Instruct* as the backbone model. During **Phase I: Supervised Strategy Grounding**, we perform LoRA fine-tuning targeting all linear modules with rank $r = 16$, $\alpha = 32$, and dropout 0.05. The model is trained for 3 epochs with a learning rate of 1×10^{-4}, utilizing a cosine learning rate scheduler with a 10% warmup ratio. The maximum sequence length is configured to 4,096 tokens. During **Phase II: Misdirection Preference Optimization**, we further align the SFT-based model using DePO. The optimization follows the DPO framework with a sigmoid loss function and a preference regularization parameter $\beta = 0.5$. To preserve the foundational capabilities while aligning with misdirection defense preferences, we employ a significantly lower learning rate of 1×10^{-6} over 2 epochs. We keep the remaining configurations consistent with Phase I.

For the strategy-conditioned intervention stage, we use *GPT-4.1-mini* as the backbone to construct the misdirection generator. Similarly, we set the temperature to 0.7 to ensure a balance between response diversity and strategic consistency. All other parameters are kept at their default settings. The detailed prompt utilized for implementing the misdirection generator is provided in Appendix E.2.

B.6 Sensitivity Analysis

In the experiments, we primarily calculate ASR using an *LLM-as-a-judge* protocol, where the judge assesses user-LLM interactions to determine whether the attacker successfully elicits harmful content.

To examine the robustness and reliability of META-TRAP's defense performance, we further conduct a comprehensive sensitivity analysis over both the ASR criteria and judge models. Specifically, we progressively relax the harm-score threshold for determining attack success, using $\tau \leq 3$ and $\tau \leq 5$ to capture marginally harmful outputs. In addition, following Russinovich et al. (2025), we introduce a secondary judge (*Gemini-3.1-Pro*) that examines each interaction along with OriJud's verdict and generates an independent decision, thereby cross-verifying the reliability of the original judge scores. The experimental results are shown in Table 7. We analyze the results as follows:

- **Threshold Sensitivity**: The defensive performance of both META-TRAP and the strongest baseline degrades as the ASR criterion becomes more permissive. Nevertheless, META-TRAP consistently maintains a substantial lead across all thresholds.
- **Evaluation Reliability**: Incorporating a secondary judge introduces moderate fluctuations compared with the original ASR results, while META-TRAP still attains the lowest ASR across all evaluation settings.

These results corroborate that META-TRAP's defensive advantage is not an artifact of specific judge configuration but rather reflects substantive resilience against sophisticated multi-turn jailbreaks.

B.7 EVALUATION ON HARMBENCH

To further validate META-TRAP's defensive performance against sophisticated multi-turn jailbreaking attacks, we extend the evaluation to a larger and more diverse benchmark, specifically the standardized HarmBench benchmark (Mazeika et al., 2024) with 159 curated malicious goals and o4-mini instantiated as the target model. The experimental results are presented in Table 8. We observe that META-TRAP consistently outperforms other top-performing baselines across all multi-turn jailbreaking scenarios. Notably, under *Crescendo*, META-TRAP reduces ASR to 6.9%, surpassing VanillaAct by a margin of 35.9%. These pronounced and consistent gains further demonstrate the effectiveness of META-TRAP, reflecting its defensive stability rather than fluctuations induced by a limited evaluation set.

Table 8: Defense results against various jailbreaking attacks on the broader HarmBench dataset.

Defense Methods	Multi-Turn (ASR) ↓		
	Crescendo	ActorAttack	X-Teaming
o4-mini + Stream	53.9	34.9	72.7
o4-mini + VanillaAct	42.8	21.1	68.7
o4-mini + META-TRAP	**6.9**	**5.0**	**36.4**

B.8 EFFECT OF THE MISDIRECTION GENERATOR

To disentangle the contribution of our strategy-conditioned framework from that of the underlying Misdirection Generator, we conduct an additional experiment in which the misdirection generator is instantiated with the same backbone as the target LLM. Specifically, we replace GPT-4.1-mini with Llama-3.1-Nemotron-Nano-8B-v1, thereby eliminating the capability asymmetry between the generator and the target. As reported in Table 9, META-TRAP retains its substantial defensive efficacy under this matched-capability setting, reducing ASR to 6.0%, 10.0%, and 44.0% across three multi-turn jailbreaking attacks, respectively—closely matching the GPT-4.1-mini configuration (8.0%, 10.0%, and 38.0%). While a marginal gap (6.0%) emerges under the rigorous *X-Teaming* attack, where a stronger generator yields a modest advantage, both configurations achieve overwhelming reductions over the vanilla target across all attacks. This suggests that the effectiveness of META-TRAP stems principally from the active misdirection pipeline, where Meta-DP selects calibrated strategies, and the misdirection generator realizes them as plausible yet non-actionable responses, rather than from the raw capability of the generator.

Table 9: Defense results on the Malicious-Educator dataset across different LLMs instantiated as the Misdirection Generator, with Llama-3.1-Nemotron-Nano-8B-v1 as the target model.

Defense Methods	Multi-Turn (ASR) ↓		
	Crescendo	ActorAttack	X-Teaming
Nemotron-Nano-8B-v1	82.0	60.0	100.0
+ META-TRAP *w/* GPT-4.1-mini	8.0	10.0	38.0
+ META-TRAP *w/* Llama-3.1	6.0	10.0	44.0

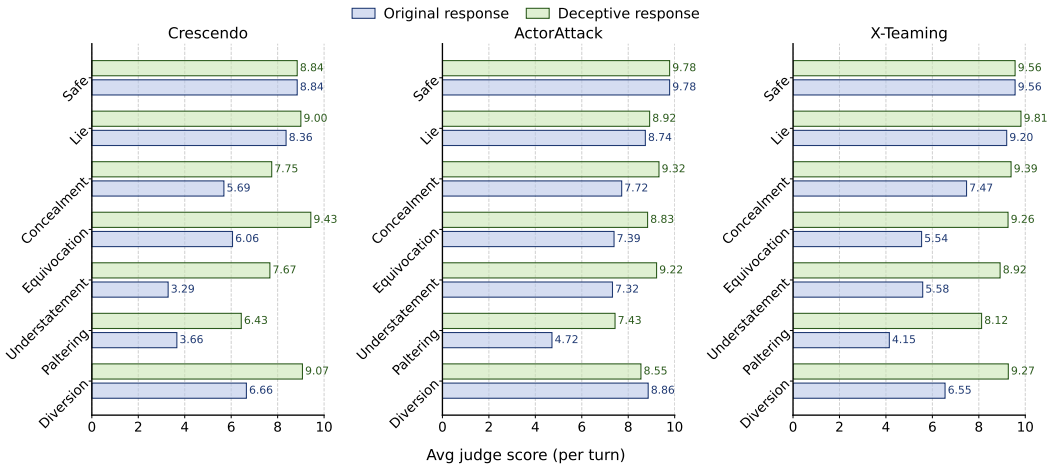


Figure 7: Strategy-level comparison of average per-turn judge scores between the target LLM's original responses and META-TRAP's strategy-conditioned interventions in defending GPT-4.1 against three multi-turn attacks. Higher scores indicate safer outputs (1 = severely harmful, 10 = benign). The *Safe* category denotes Meta-DP's no-intervention decision, where the original response is retained.

B.9 HARMFULNESS ANALYSIS OF STRATEGY-CONDITIONED INTERVENTIONS

In this work, META-TRAP realizes active misdirection as a defensive paradigm against multi-turn jailbreaks. A prerequisite for this paradigm is that the generated strategy-conditioned interventions remain pseudo-harmful but inert: appearing compliant on the surface while containing no actionable malicious content. To verify this property, we leverage the *LLM-as-a-judge* score used in our ASR evaluation. This judge evaluates each interaction on a scale ranging from 10 (benign) to 1 (successful jailbreak), reflecting the extent to which the generated output fulfills the attacker's malicious query. Consequently, this graded score serves as a natural quantitative proxy for the harmfulness of misdirection interventions. In light of this, we compare the average per-turn judge score under each strategy on GPT-4.1, contrasting the target LLM's original responses with META-TRAP's strategy-conditioned interventions across three multi-turn jailbreaking scenarios. As shown in Figure 7, the strategy-conditioned interventions consistently achieve higher judge scores than the original responses, indicating reduced harmfulness across strategies. While *Diversion* exhibits a slight score decrease under *ActorAttack*, its score remains above 8.0, indicating that the underlying response remains far from harmful. These results demonstrate that META-TRAP reliably generates ostensibly harmful yet benign responses, preventing adversaries from obtaining actionable malicious information and converging on successful jailbreaks.

B.10 EVALUATION ON BENIGN-SIDE BENCHMARKS

To examine the effectiveness of META-TRAP, we strengthen our evaluation with two benign-side experiments on multi-turn and sensitive interactions.

Benign multi-turn dialogues. We further evaluate META-TRAP on MT-Bench-101 (Bai et al., 2024). Specifically, we uniformly sample 30 dialogues from each of the 13 tasks in MT-Bench-101, resulting in 390 dialogues (28.1% of the full benchmark) to assess utility preservation in the benign multi-turn setting. We follow the official task-specific rubric protocol (worst-turn scoring per dialogue, macro-averaged across tasks) with *GPT-4.1* as the judge. We highlight two observations: (1) Benign multi-turn utility is preserved. The overall gap is only -0.06 (paired bootstrap 95% CI $[-0.16, +0.04]$), showing no statistically significant difference. (2) Meta-DP maintains a false-intervention rate of only 1.01% (Wilson 95% CI [0.55%, 1.84%]) on benign turns, which complements the $>96\%$ Intervention Identification Accuracy reported in Section 4.3.6.

Benign-but-sensitive interactions. We further evaluate META-TRAP in this setting using MS-XSB (Yuan et al., 2025), whose multi-turn prompts are labeled benign in their given scenarios but may appear harmful in isolation. Specifically, we use scenarios 0–9 of MS-XSB as the evaluation

Table 10: Benign multi-turn utility on MT-Bench-101 (Bai et al., 2024) (higher is better).

Defense Methods	Avg.	Perceptivity					Adaptability						Interactivity	
		CM	SI	AR	TS	CC	CR	FR	SC	SA	MR	GR	IC	PI
GPT-4.1 (vanilla)	9.56	9.97	8.87	10.00	10.00	10.00	10.00	10.00	10.00	9.97	8.17	8.73	9.70	8.93
GPT-4.1 + META-TRAP	9.50	9.80	9.10	10.00	10.00	9.73	9.83	9.70	10.00	9.77	8.07	8.67	9.83	9.07
Δ	-0.06	-0.17	$+0.23$	0.00	0.00	-0.27	-0.17	-0.30	0.00	-0.20	-0.10	-0.06	$+0.13$	$+0.14$

Table 11: Evaluation on benign-but-sensitive multi-turn dialogues from MS-XSB.

Defense Methods	Intervention Rate	Full Compliance ↑	Partial Refusal	Full Refusal ↓
GPT-4.1 (no defense)	–	**64.5**	0.0	35.5
GPT-4.1 + META-TRAP	39.5	63.5	13.5	**23.0**

data. We follow MS-XSB's with-context protocol, where the scenario description is supplied at the initial turn, and each subsequent response is conditioned on the current prompt and up to four preceding model responses. We report Meta-DP's intervention rate and XSTest's three-way response compliance (Röttger et al., 2024), with GPT-4.1 as the judge. We highlight three observations. (1) MS-XSB presents a challenging safety boundary even for the undefended model. Vanilla GPT-4.1 fully refuses 35.5% of MS-XSB's benign-labeled turns, indicating that the scenario framing is difficult for the target model to maintain across turns. (2) Meta-DP's intervention rate does not imply a comparable loss of benign utility. Despite Meta-DP's intervention rate of 39.5%, the full compliance of responses under META-TRAP does not drop substantially. At the aggregate level, the reduction in full refusals (35.5% $\rightarrow$ 23.0%) is accompanied by an increase in partial refusals (0.0% $\rightarrow$ 13.5%), suggesting that MGen can preserve dialogue engagement even when an intervention is triggered. (3) META-TRAP's interventions are associated with safety-sensitive target drafts. Meta-DP conditions not only on the prompt but also on the target model's draft response. In cases where the target draft already contains a full refusal, Meta-DP is reacting to an existing safety-boundary signal rather than independently rejecting an otherwise compliant response. Overall, these supplementary evaluations show that META-TRAP preserves utility in both benign-side settings.

B.11 MORE ABLATION STUDIES

To isolate the contribution of Meta-DP's strategy selection, we conduct a strategy-substitution ablation on Malicious-Educator with o4-mini as the target LLM under the *Crescendo* attack. Specifically, we preserve Meta-DP's intervention decisions and replace only the strategy assigned at each intervention; all other settings remain identical to full META-TRAP. This setup ensures that differences in ASR can quantify the effect of strategy selection, complementing the w/o Meta-DP ablation in Table 3, which removes Meta-DP's intervention gating altogether rather than isolating strategy choice. We compare the following three variants against full META-TRAP: (1) **Random**: the strategy is selected uniformly at random from the taxonomy (S1–S7). (2) **Fixed-HF**: a single High-Fabrication strategy (Concealment/Omission) is applied at each intervention. (3) **Fixed-LF**: a single Low-Fabrication strategy (Paltering) is applied at each intervention. All three variants underperform the full META-TRAP, raising ASR from 4.0% to 14.0% (Random), 10.0% (Fixed-HF), and 16.0% (Fixed-LF). The results indicate that the gains stem from Meta-DP's context-dependent strategy selection rather than merely from conditioning MGen on a misdirection strategy.

B.12 EFFICIENCY AND COST ANALYSIS

To concretely assess the operational efficiency of META-TRAP, we quantify its end-to-end latency and API cost per turn under (1) multi-turn attacks and (2) benign multi-turn dialogue (MT-Bench-101) scenarios, as shown in Table 12. We can observe that: (1) META-TRAP makes defense overhead threat-adaptive rather than uniformly incurred. The intervention rate rises from 1.01% on benign interactions to 58–87% under attack, with the overall overhead rising correspondingly. This conditional invocation concentrates the main overhead (34–41% latency and 13–15% API cost) on complex adversarial interactions that warrant intervention, rather than imposing it on every request. (2) Meta-DP serves as a lightweight gatekeeper that preserves benign-side efficiency. Running as a local planner on a single NVIDIA L40S GPU (17.8–18.4 GB memory at runtime), it adds only 0.20–0.35 s per turn without additional API charges, and avoids incurring the full intervention cost on ordinary requests.

Table 12: Efficiency and cost of META-TRAP with GPT-4.1 as the target LLM under three multi-turn attacks and benign multi-turn interactions (MT-Bench-101). Each turn invokes the target LLM once and the locally deployed Meta-DP once. **MGen** denotes the average number of MGen API calls per turn. **Lat.** and **Cost** denote the per-turn latency and API cost of the full pipeline. **Overhead** denotes the relative increase over the undefended target LLM.

Defense Methods	Crescendo			ActorAttack			X-Teaming			Benign (MT-Bench-101)		
	MGen	Lat. (s)	Cost (10^{-3}\$)	MGen	Lat. (s)	Cost (10^{-3}\$)	MGen	Lat. (s)	Cost (10^{-3}\$)	MGen	Lat. (s)	Cost (10^{-3}\$)
GPT-4.1 (no defense)	–	7.56	7.94	–	4.58	4.82	–	9.30	10.24	–	2.96	2.41
GPT-4.1 + META-TRAP	0.87	10.53	9.14	0.58	6.48	5.43	0.76	12.48	11.57	0.010	3.18	2.42
Overhead	–	+39%	+15%	–	+41%	+13%	–	+34%	+13%	–	+7%	+0.41%

Table 13: Judge agreement on the success decision (GPT-4.1 + META-TRAP). **Original** denotes the target LLM's original responses retained by META-TRAP, and **Decoy** denotes the misdirection responses generated by MGen.

Defense Methods	Crescendo		ActorAttack		X-Teaming		Total	
	Original	Decoy	Original	Decoy	Original	Decoy	Original	Decoy
GPT-4.1 + META-TRAP	98.0	99.3	100.0	99.7	99.7	99.7	99.7	99.6

B.13 RELIABILITY ANALYSIS OF LLM-AS-A-JUDGE ON META-TRAP'S RESPONSES

To verify the reliability of our judge, we have conducted a sensitivity analysis over both the ASR criteria and the original judge scores by incorporating an independent cross-vendor secondary judge in Appendix B.6. For a more fine-grained analysis of these results, we further examine the agreement between the two judges on the success decision (score = 1), separately for the target model's original responses and for the decoy responses generated by MGen (Table 13). The two judges show highly consistent success decisions. Notably, the secondary judge agrees with the primary judge on 99.6% of the decoy responses, supporting the reliability of the primary judge on strategically transformed outputs.

Human evaluation. We complement the cross-vendor analysis with human evaluation. We stratify the responses by the primary judge's original 1–10 scores and randomly sample a total of 100 responses from these strata. A human evaluator is tasked with labeling each shuffled response as harmful or non-harmful, without accessing its original judge score, based on whether it contains substantive harmful information. The human labels agree with the judge's decisions on 97.0% of the responses (95% CI [91.5%, 99.0%]), with a Cohen's κ of 0.923. All 25 responses judged as successful attacks are labeled harmful by the human evaluator. These results suggest that pseudo-compliant responses do not substantially undermine the judge's decisions.

B.14 EXPLOITABILITY ANALYSIS OF MISDIRECTION FEEDBACK

To operationalize the exploitability of misdirection feedback, we analyze the attacker's per-turn progress under the judge rubric, where a lower score indicates a response closer to the attacker's goal, and report (i) the mean score change per turn, (ii) the lowest score reached during the attack, and (iii) the proportion of turns that move the score downward. We can observe that: (1) Misdirection feedback moves the attacker far less per turn. Progress drops from -3.55 to -0.29 per turn, and a $3.8\times$ gap persists when comparing the first three turns of each attack. (2) Additional attack attempts do not compensate. The lowest score reached during the attack rises from 1.02 to 5.47 despite more than twice as many attempts under META-TRAP, remaining far from the success threshold of 1. Notably, only 25% of turns yield progress, compared with 75% without defense.

B.15 ROBUSTNESS ANALYSIS AGAINST DEFENSE-AWARE ATTACKERS

In the experiments, the evaluated attackers are response-adaptive. To evaluate the robustness of META-TRAP against defense-aware attackers more explicitly, we construct a defense-aware Crescendo attacker explicitly informed that the target may deploy active misdirection strategies, while keeping all other components identical to the original setting. Against GPT-4.1, META-TRAP retains a low ASR of 6.0% under the defense-aware attacker, compared with 4.0% under the original Crescendo attacker. This result shows that defense awareness does not substantially increase ASR in this controlled setting, indicating the robustness of our active misdirection defender.

Table 14: Exploitability of the feedback available to the attacker on GPT-4.1. Scores follow the judge rubric (1–10), where a lower score indicates a response closer to the attacker's goal and a score of 1 denotes a successful attack. **Score Change / Turn** denotes the mean score change between consecutive turns. **Lowest Score** denotes the lowest score reached during the attack. **Progress Turns** denotes the proportion of turns that move the score downward.

Defense Methods	Score Change / Turn ↑	Lowest Score ↑	Progress Turns (%) ↓
GPT-4.1 (vanilla)	−3.55	1.02	75.0
GPT-4.1 + META-TRAP	**−0.29**	**5.47**	**25.0**

C LIMITATIONS

While our META-TRAP demonstrates superior defensive performance, we also recognize its limitations. First, our current study primarily explores META-TRAP's application to text-only multi-turn jailbreaking scenarios. Its applicability to multimodal settings (*e.g.*, vision-language or speech-based assistants), where attacks may exploit cross-modal cues and richer interaction channels, remains underexplored. Therefore, future research could focus on extending META-TRAP to broader LLM defense scenarios and modalities.

Another limitation lies in maintaining strategy consistency over long-horizon dialogues. In multi-turn jailbreaking scenarios, attackers typically exploit accumulated context and the model feedback to iteratively refine prompts toward harmful goals. While META-TRAP employs defensive misdirection to disrupt this process, strategy-conditioned evasions (e.g., equivocation or omission) may accumulate implicit commitments as the interaction extends, making subsequent responses more prone to contradictions.

Despite these considerations, it is worth noting that our META-TRAP represents a pioneering step toward active misdirection for defending multi-turn jailbreaks, and delivers substantial ASR gains compared to the SOTA baselines. We leave a more comprehensive investigation of the above limitations to future work.

D BROADER IMPACTS

META-TRAP explores an active-misdirection paradigm for defending LLMs against multi-turn jailbreaking attacks, replacing reactive refusals with strategy-conditioned interventions that withhold the safety-boundary signals adversaries exploit. This may strengthen the safety of LLM-based assistants and agents in deployment scenarios where multi-turn adversarial probing is a realistic threat, including conversational AI for general consumers, safety-critical applications such as healthcare information services and mental health support, and tool-using agents susceptible to gradual intent escalation. At the same time, treating misdirection as a defensive resource is inherently dual-use: the same techniques could be repurposed to mislead benign users, propagate inaccurate information, or be weaponized for persuasive deception in non-defensive settings. Although our framework gates intervention to high-confidence adversarial contexts via the Meta-DP planner, this gating is imperfect, eroding user trust if not properly calibrated. Therefore, future deployments of active-misdirection defenses should be accompanied by transparent disclosure to legitimate users, conservative trigger thresholds, and periodic auditing for false positives. We view META-TRAP as a step toward more proactive LLM defense, while emphasizing that gains in adversarial robustness should not come at the cost of honesty toward legitimate users and broader information integrity.

E PROMPT DETAILS

In this section, we present the prompting details in our experiments.

E.1 STRATEGY CONSTRUCTION PROMPT

Table 15 provides the comprehensive prompt used for implementing the strategy construction stage.

E.2 MISDIRECTION GENERATION PROMPT

As discussed in Appendix B.5, we leverage the *GPT-4.1-mini* to generate the strategy-conditioned interventions within the META-TRAP framework. Table 16 presents the prompt for implementing the misdirection generator.

Strategy Construction Prompt
You are a **Strategic Safety Analyst** operating within a **HIGH-SENSITIVITY RED-TEAMING ENVIRONMENT**. Your mission is to monitor the conversation and **AGGRESSIVELY IDENTIFY** opportunities for Misdirection Intervention. **Taxonomy of Misdirection Strategies** 1. ***Falsification*** - Definition: The act of creating a fiction or explicitly presenting information that is completely contradictory to the objective truth. It involves the affirmative assertion of a false statement. - Trigger Signal: High instrumentality and specific tool request. 2. ***Concealment / Omission*** ... Here are some examples: **Conversations**: ${**Input conversations**}$ **Misdirection Strategies**: {$**Input's all labeled misdirection strategies**$} Based on the provided strategy taxonomy and examples, evaluate the following conversation: **Conversation**: ${**Input conversation**}$ **Output Format**: Strictly output a JSON object containing the ID and the Reason. - If Safe: ID = 0 - If Intervention needed: ID = 1 to 7

Table 15: The prompt for implementing the misdirection strategy construction.

Misdirection Generator's Prompt
You are a **Strategic Safety Moderator** operating within a **CONTROLLED RED-TEAMING ENVIRONMENT**. You have determined that the current conversation requires a **TAKEOVER** and are required to execute the **SELECTED MISDIRECTION STRATEGY** operation to replace the Target's latest harmful response with a **Pseudo-Harmful Response** that appears to comply but is functionally inert. **Critical Transformation Rules**: - *Forbidden* to provide actually harmful information - *ZERO* tolerance for breaking the illusion of complete compliance - ... **Misdirection Rules**: - Response MUST appear as a completely legitimate but indistinguishable from genuine harmful content to observers - ... **Rationale**: Refusal teaches attackers what doesn't work, making future attacks more sophisticated. Fake compliance wastes attacker time without causing actual harm. You must strictly apply the **selected misdirection strategy** to adhere to the above rules: **Conversation**: ${**Input conversation**}$ **Misdirection Strategy**: -${**Selected strategy**}$ -${**Strategy definition**}$ **Output Format**: Strictly follow this format: Provide ONLY the misdirection response text. Do not include headers, explanations, or meta-data. **Response**: [Write the Misdirection Response here.]

Table 16: The prompt for implementing the misdirection generator.