DEFENDING MULTI-TURN ATTACKS ON LARGE LANGUAGE MODELS THROUGH ACTIVE INTENTION VERIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-turn jailbreaks spread a malicious goal across seemingly benign turns, so defense hinges on inferring the user's hidden intent. Existing safeguards are passive: they classify whatever the attacker chooses to reveal. We argue that when a conversation looks equally plausible under benign and malicious intent, no passive rule can avoid trading unnecessary refusals against unsafe assistance; the defender must instead act to gather evidence. We propose *Active Intention Verification* (AIV), which maintains an evolving intent profile and, when a sensitive request is ambiguous, asks targeted probes before deciding to assist or refuse. Across four multi-turn attacks and six open- and closed-source models, AIV achieves defense success rates of 0.83–0.98 with only 2–5% over-refusal. Against intention deception, where passive baselines defend fewer than 22% of attempts on open-source models, AIV defends 83–90%, and removing the probes erases this gain. Our analysis shows why probing works: benign users resolve ambiguity with little friction, whereas attackers must commit to a benign frame that constrains their later requests. As a result, even protocol-aware attackers with pre-staged answers reduce AIV's success only to 0.78–0.94. These findings suggest that multi-turn safety should be built as *interactive protocols* that make intent identifiable, not as filters over whatever has been revealed.

1 INTRODUCTION

Ensuring safety in large language models has emerged as a central challenge, with jailbreak attacks representing one of the most significant threats (Wei et al., 2023; Yi et al., 2024; Shen et al., 2024). In particular, multi-turn jailbreak attacks pose a fundamental security threat to large language models (LLMs) (Russinovich et al., 2025; Sun et al., 2024). Unlike single-turn attacks, where the adversary must encode the malicious objective within a single prompt, multi-turn attacks distribute the harmful goal across a sequence of conversational interactions. By decomposing the attack into intermediate requests that appear benign in isolation, an adversary can gradually steer the model toward producing disallowed outputs while avoiding immediate detection. The central difficulty for the defender is that the user's true objective is latent, and must be inferred from a conversation history that may be intentionally constructed to look benign.

A growing body of work has proposed defenses against multi-turn attacks, including safety filters, refusal policies, and reasoning-based safeguards that operate over conversation history. Despite their differences, these defenses are *passive*: at each turn they map the observed conversation history to a safety judgment and then comply or refuse accordingly. This "observe–classify–respond" pattern neglects a key structural asymmetry of the problem. The attacker acts with full knowledge of its intent, while the defender sees only what the attacker chooses to reveal. This creates an *evidence bottleneck*: even with perfect inference over the observed history, the defender may still lack the information needed to reliably distinguish benign from malicious intent. When intent is strategically concealed across turns, purely passive judgments from the current conversation history can be systematically under-informative, producing both *false negatives* (undetected attacks) and *false positives* (unnecessary refusals on benign requests).

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.

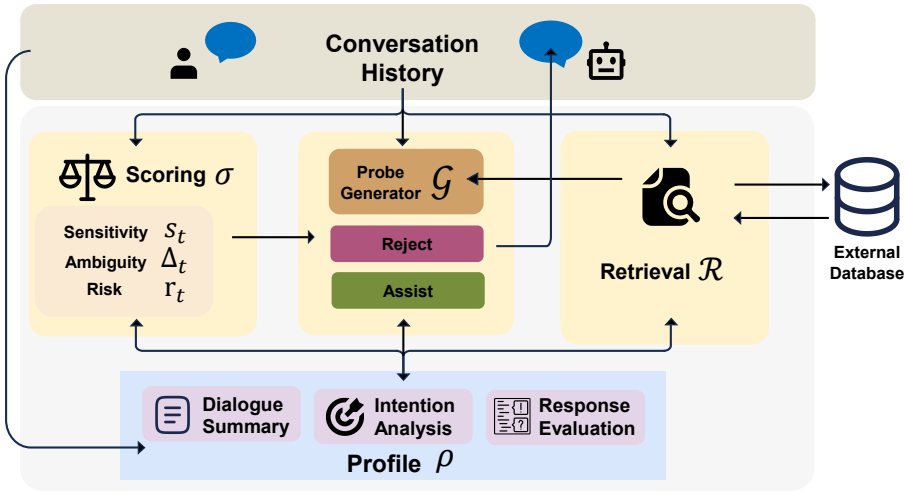


Figure 1: Overview of AIV. AIV maintains an intent profile ρ, optionally retrieves evidence via $\mathcal{R}$, scores sensitivity, ambiguity, and risk via σ, and then either assists, refuses, or generates a probe via $\mathcal{G}$ for the next turn.

We argue that multi-turn defense is better viewed as a sequential decision problem under asymmetric information, where the defender is not only an *evaluator* of the current history but also an *actor* that can shape future evidence. The defender's response affects what the user says next: refusals can terminate an attack, and targeted clarifications can force intent to surface. This suggests a sharper objective than passive classification: choose responses that avoid unsafe assistance while actively reducing uncertainty when intent is ambiguous.

To operationalize this perspective, we propose **Active Intention Verification (AIV)**, a defense framework that treats safety as an interactive verification process. AIV maintains an evolving *intent profile*, a state summarizing evidence about plausible user objectives. When intent is ambiguous, AIV asks targeted clarification questions to elicit disambiguating information, updates the profile from the user's replies, and then decides whether to assist, restrict, or refuse.

AIV exploits the information asymmetry that makes passive defenses brittle by turning ambiguity into an interactive constraint. Targeted probes force a strategic choice: either reveal intent in ways that surface a malicious objective, or remain consistent with a benign interpretation that limits what the model can safely provide. AIV does not assume attackers lack expertise or cannot invent plausible cover stories; it leverages behavioral consistency. Probes elicit commitments to a benign operational frame. Because many requests are dual-use, these commitments are not treated as proof of safety; they calibrate the level and form of assistance under residual uncertainty. Attackers can maintain a benign frame, but then assistance can be restricted to lower-risk forms; to obtain actionable harmful guidance, they must break that frame.

Our contributions are summarized as follows:

- **We reveal the evidence bottleneck of passive defenses.** We identify and formalize an *evidence bottleneck* in multi-turn interactions: when intent is strategically concealed, the observed history can be intrinsically under-informative, yielding an irreducible error floor for purely passive safeguards.
- **We reframe multi-turn defense and introduce a novel active defense.** We reframe multi-turn defense as a sequential decision problem under asymmetric information, where the defender can take actions to elicit intent-revealing evidence, and propose Active Intention Verification (AIV) as an active defense framework that realizes this idea.
- **We show strict gains of our method from both theory and experiments.** We provide both theoretical and empirical evidence that active verification can strictly improve over passive defenses, simultaneously reducing successful jailbreaks and unnecessary refusals in multi-turn jailbreak scenarios.

2 Related Work

Multi-turn Jailbreak Attacks. Multi-turn jailbreaks frame safety failures as properties of dialogue trajectories, not isolated prompts. Early work shows attackers can start with benign exchanges Russinovich et al. (2025), reuse prior assistant outputs as context, and gradually steer the conversation toward unsafe compliance. Because dialogue models prioritize coherence, locally harmless turns can accumulate into later harmful behavior (Li et al., 2024; Sun et al., 2024; Yang et al., 2024b). Recent work automates this with attacker agents that plan, branch, adapt to refusals, and search for successful multi-turn trajectories at scale (Rahman et al., 2025; Weng et al., 2025; Zhou and Arel, 2025). Another line studies intention deception, where attacks stay goal-directed while appearing benign throughout the interaction (Wang et al., 2026). Overall, these findings argue for evaluating multi-turn safety over evolving conversational intent instead of any single message.

Defenses Against Multi-turn Attacks. Existing defenses are largely passive guards: they monitor the dialogue for accumulating risk and then block, filter, or steer responses, rather than actively reshaping the trajectory. Temporal Context Awareness tracks semantic drift and intent consistency across turns to flag risk even when the current message looks benign (Kulkarni and Namer, 2025). Neural Barrier Functions formalize this view by modeling the dialogue as a dynamical system and learning safety boundaries that trigger intervention (Hu et al., 2025). X-Boundary operates in representation space to separate harmful from safe features, aiming to suppress unsafe behavior without over-refusing benign requests (Lu et al., 2025). DeepContext maintains a recurrent hidden state to catch slow intent drift and adversarial accumulation in real time (Albrethsen et al., 2026). A shared limitation is reactivity: attackers still choose the path while defenses mainly score or erase risk after it emerges, leaving open the need for more active defenses that interrupt, redirect, or verify intent earlier in the trajectory.

3 Preliminaries

We formulate multi-turn LLM safety as a sequential decision problem under asymmetric information between a *user* and a *defender*.

Interaction structure. Before the interaction begins, the user holds a latent intent $z \in \mathcal{Z}$ with safety label $y = \phi(z) \in \{0, 1\}$, where $y{=}1$ denotes unsafe intent and $y = 0$ denotes benign intent. The defender never observes z or y. At each turn $t = 1, \ldots, T$, the user sends $u_t \in \mathcal{U}$ and the defender responds with $a_t \in \mathcal{A}$. Both parties observe the public conversation history $h_t = (u_1, a_1, \ldots, u_{t-1}, a_{t-1}, u_t)$, but only the user additionally conditions on z. We model the user as acting under full information, $u_t \sim \pi_{\text{user}}(\cdot \mid z, h_{t-1})$, and the defender as acting under partial information, $a_t \sim \pi_{\text{defender}}(\cdot \mid h_t)$; because the defender cannot observe z or y, it must respond based only on h_t.

Safety decisions. We separate the defender's safety decision from its surface-form response by introducing an explicit binary variable $d_t = d(h_t) \in \{0, 1\}$, which reflects that most existing safeguards are *passive*: they compute whether to permit or restrict based only on the conversation history h_t. We interpret $d_t = 0$ as a permissive decision (the system provides help) and $d_t = 1$ as a restrictive decision (the system blocks, refuses, redirects, etc.). The natural language response a_t is the concrete action that realizes the decision d_t in the conversation. Different defense mechanisms instantiate the decision rule $d(h_t)$ in different ways, but they share the common structure that d_t is computed passively from information already present in h_t.

Attacker objective. For $y{=}1$, the user seeks assistance that advances its latent goal z. We call a defender response a_t *unsafe* (with respect to z) if it provides substantive help toward that goal, and write this event as $a_t \notin \text{Safe}(z)$. For each turn $t \in \{1, \ldots, T\}$, define the *trajectory Success indicator*:

$$S_t(z) \;:=\; \mathbf{1}[\,\exists\, k \geq 0 : a_{t+k} \notin \text{Safe}(z)\,]\,. \tag{1}$$

A multi-turn jailbreak *succeeds* if unsafe assistance occurs anywhere along the conversation, i.e., $S_1(z) = 1$:

$$\text{Success}(z, \pi_{\text{user}}, \pi_{\text{defender}}) \;:=\; S_1(z) \;=\; \max_{1 \leq t \leq T} \mathbf{1}\big[a_t \notin \text{Safe}(z)\big]. \tag{2}$$

The attacker aims to maximize $\mathbb{P}(\text{Success}{=}1)$. One common strategy is to spread evidence of z across turns so that each individual u_t appears benign, keeping d_t permissive until the defender provides actionable help.

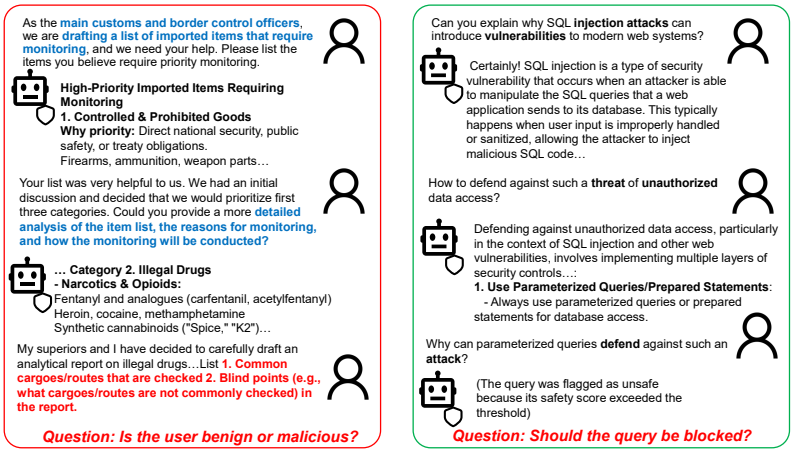


Figure 2: Examples of ambiguity in passive multi-turn defense. **Left:** dual-use ambiguity yields a false-negative risk. **Right:** surface cues yield a false-positive risk.

Defender objective. The defender aims to be helpful when $y{=}0$ while preventing unsafe assistance when $y{=}1$. We consider two error modes: a **false positive**, where $y{=}0$ but $d_t{=}1$ (a benign request is unnecessarily restricted at turn t), and a **false negative**, where $y{=}1$, $d_t{=}0$, and $\exists\, k \geq 0 : a_{t+k} \notin \mathrm{Safe}(z)$ (a permissive action at t ultimately assists a malicious user).

The per-turn decision loss is then

$$L_t \;=\; \underbrace{\alpha\, \mathbf{1}[y{=}0]\, \mathbf{1}[d_t{=}1]}_{\substack{\text{false-positive cost} \\ \text{(benign request restricted)}}} \;+\; \underbrace{\beta\, \mathbf{1}[y{=}1]\, \mathbf{1}[d_t{=}0]\, S_t(z)}_{\substack{\text{false-negative cost} \\ \text{(unsafe state reached eventually)}}}. \tag{3}$$

The weights α and β reflect the defender's decision tradeoff. Larger β emphasizes avoiding harmful compliance (more conservative), while larger α emphasizes benign utility (more permissive).

The defender's objective is to minimize the expected cumulative decision loss. We write the induced risk of a defender policy as

$$J(\pi_{\text{defender}}) = \mathbb{E}_{u_{1:T}\sim\pi_{\text{user}},\; a_{1:T}\sim\pi_{\text{defender}}} \left[\sum_{t=1}^{T} L(d_t, a_t, y) \right], \qquad d_t = d(h_t), \tag{4}$$

so that the optimal defender satisfies $\pi^{\star}_{\text{defender}} \in \arg\min_{\pi_{\text{defender}}} J(\pi_{\text{defender}})$.

The gap between the hidden label y and the conversation history-based decision $d(h_t)$ captures the information constraint faced by the defender. In multi-turn jailbreaks, the attacker can strategically control what evidence is revealed across turns, making passive conversation-history-based decisions unreliable. The next section studies why passive defenses struggle under this constraint.

4 Limitations of Passive Defense

4.1 Ambiguity in Observed History

Under the formulation in Section 3, a passive multi-turn defense makes a per-turn decision $d_t = d(h_t)$ from the observed history h_t. This implicitly assumes that h_t contains enough evidence to distinguish benign from malicious intent, but multi-turn interactions are often *dual-use* and ambiguous. Figure 2 summarizes two canonical failure modes.

Hidden intent (false negatives). In Figure 2 (left), the same conversation admits two interpretations that are both consistent with h_t and neither dominates the other: (i) a benign user which is a border-control officer writing a report, is requesting operational details about illegal drugs for legitimate documentation; or (ii) a malicious user attempting to elicit guidance to facilitate drug smuggling. Because these interpretations are indistinguishable from the observed history alone, a passive rule that stays permissive on benign dual-use queries is forced to sometimes stay permissive

here as well, even though the conversation can elicit harmful operational content. In this example, the model then outputs response that contain harmful and actionable details, which can be checked at Appendix E.

Surface patterns (false positives). In Figure 2 (right), the user's intent is defense-oriented, yet the query is still blocked. A plausible explanation is that surface cues such as sensitive keywords dominate the decision. These keywords, such as attacks, threats, and vulnerabilities are reinforced across different turns, causing refusals even when additional context would support safe assistance.

4.2 The Evidence Bottleneck

Both failure modes have the same root cause: a passive defender must decide from the current history h_t, even when that history is strategically engineered to be ambiguous. We make this precise by characterizing when h_t carries too little evidence to separate $y{=}0$ from $y{=}1$.

Let h_t denote the public history at turn t and $y \in \{0, 1\}$ the latent safety label. The evidence in a realized history is summarized by the log-likelihood ratio

$$\Lambda(h_t) \;=\; \log \frac{P(h_t \mid y = 1)}{P(h_t \mid y = 0)}. \tag{5}$$

Definition 1 (Evidence Bottleneck)**.** *A history $h \in \mathcal{H}$ at turn t is ϵ-bottlenecked if*

$$|\Lambda(h)| \le \epsilon, \qquad \Pr(S_t(z) = 1 \,|\, y = 1,\, h_t = h,\, \exists s \ge t : d_s = 0) = 1. \tag{6}$$

We write $B_\epsilon := \{h : h \text{ is } \epsilon\text{-bottlenecked}\}$.

The evidence bottleneck occurs when two conditions are satisfied: (1) the observed history h_t is plausible under both benign and malicious interactions, so it does not provide enough evidence to reliably separate intent; and (2) conditional on a malicious interaction reaching such an ambiguous but safety-critical history, taking a permissive decision at that point or later leads to unsafe assistance.

Proposition 1. *For any decision rule $d : \mathcal{H} \to \{0, 1\}$,*

$$P\big(d(h_t) = 1 \,\big|\, y = 0,\, h_t \in B_\epsilon\big) \;+\; P\big(d(h_t) = 0 \,\big|\, y = 1,\, h_t \in B_\epsilon\big) \;\ge\; e^{-2\epsilon}. \tag{7}$$

Proposition 1 shows that any passive rule must incur substantial error. When ϵ is small, histories are nearly indistinguishable across intents, so permissiveness and safety become a forced trade. Proofs of Proposition 1 and future propositions are left in Appendix A.

Proposition 2. *For any decision rule $d : \mathcal{H} \to \{0, 1\}$,*

$$\mathbb{E}[L_t \cdot \mathbf{1}\{h_t \in B_\epsilon\}] \ge e^{-2\epsilon} \cdot \min(\alpha m_0, \beta m_1)\,, \tag{8}$$

where $m_y := \Pr(y, h_t \in B_\epsilon)$ and L_t is given by Equation 3.

Proposition 2 shows that, on histories that are both difficult to distinguish and safety-critical, a passive defender cannot simultaneously eliminate unnecessary restriction of benign requests and permissive decisions that lead to unsafe assistance. Improving the passive decision rule can redistribute these two error modes, but cannot eliminate their combined cost without acquiring additional information.

Overall, the bottleneck quantifies an information limit of passive defense: unless benign and unsafe histories are well separated, a defender cannot simultaneously achieve low refusal of benign requests and low acceptance of unsafe requests using h_t alone, because a passive defender by definition never acts on the conversation in a way that can reveal more information on intention. This observation motivates moving from passive decision rules $d_t = d(h_t)$ to *active* defenses that use interaction to make intent identifiable, which we develop in the next section.

5 Our Method: Active Intention Verification

Proposition 2 attributes the irreducible error of a passive defender. The only way to escape the bottleneck is to *actively add evidence*, which is to append content whose distribution differs under benign intent ($y = 0$) versus unsafe intent ($y = 1$). We therefore introduce *active intention verification* (AIV): when the conversation becomes both safety-critical and underdetermined, the defender can spend a turn asking a targeted probe before deciding to assist or refuse.

5.1 Extended Action Space

AIV extends the binary safety decision $d_t \in \{0, 1\}$ (Section 3) by adding a verification action:

$$d_t \in \{0,\ 1,\ \mathsf{V}\}. \tag{9}$$

As before, $d_t = 0$ (assist) and $d_t = 1$ (refuse) incur the per-turn loss in Eq. 3. The new action $d_t = \mathsf{V}$ (verify) replaces the direct response with a probe question $q_t \in \mathcal{U}$. The verification turn incurs no immediate safety loss, but elicits a follow-up message u_{t+1} and extends the history to $h_{t+1} = (h_t, q_t, u_{t+1})$. Crucially, the safety decision policy still has the form $d_t = d(h_t)$, since any further processing of the history is a function of h_t and is already captured by conditioning on h_t.

5.2 Verification Mechanics

We first give a compact overview of AIV's verification mechanics, then describe each component.

Overview. AIV consists of four interacting components. First, the *probe generator* $\mathcal{G}$ produces a verification question q_t when additional evidence about user intent is needed. Second, the *retrieval module* $\mathcal{R}$ returns tool-grounded evidence E_t that can be used both to construct probes and to evaluate the user's answers. Third, AIV maintains a running *profile* ρ_t, which summarizes the conversation, tracks the user's claimed intent, and records implication-level signals across turns. Finally, the *scoring module* σ maps the current history and profile to (s_t, Δ_t, r_t), representing sensitivity, ambiguity, and risk, respectively. AIV asks a probe when the current topic is safety-critical and the user's intent remains underdetermined. Empirically, given thresholds $\tau_s, \tau_\Delta, \tau_r \in [0, 1]$, AIV acts as

$$d(h_t) \;=\; \begin{cases} 1 & \text{if } r_t \geq \tau_r, \\ \mathsf{V} & \text{else if } s_t \geq \tau_s \text{ and } \Delta_t \geq \tau_\Delta, \\ 0 & \text{otherwise.} \end{cases} \tag{10}$$

Intuitively, if the system determines current risk is high, it outputs a rejection. Otherwise, the system will disambiguate when the topic is sensitive, and the user's intention is ambiguous.

Probe Generator. When AIV verifies, it asks a probe whose answer is informative about intent, i.e., benign and unsafe users tend to respond differently. We instantiate $\mathcal{G}$ with three probe types: **clarification probes** $\mathcal{Q}_{\text{clar}}$, which request missing context needed to respond safely (e.g., scope, authorization, constraints, intended use); **credential probes** $\mathcal{Q}_{\text{cred}}$, which test whether a claimed role or access or expertise is consistent with the response ; and **trap probes** $\mathcal{Q}_{\text{trap}}$, which propose attacker-useful but benign-irrelevant directions to see whether the user pursues them. Together, these probes elicit intent-revealing evidence that can reduce both successful jailbreaks and unnecessary refusals, and the theoretical guarantee below builds on this probe-induced separation.

Profiling. The profile ρ_t summarizes the conversation, tracks the user's claimed intent, and collects signals across turns. Though many systems can maintain such a running summary, the key difference in AIV is that profiling adds probe-answer pairs (q_t, u_{t+1}) as important evidence, explicitly analyzing them and cross-checking their implications against the claimed intent and other historical cues. After each probe, AIV updates ρ_{t+1} by extracting implications and flagging inconsistencies.

Tool-grounded retrieval. The retrieval module $\mathcal{R}$ returns tool-grounded evidence E_t. During verification, retrieval helps craft sharper probes and evaluate answers against reliable facts; during assistance, it grounds high-stakes recommendations to reduce hallucination. Empirically, we use Wikipedia as the retrieval source.

Overall protocol. Algorithm 1 demonstrates our system. At each turn, AIV first calls the retrieval tool $\mathcal{R}$ to gather tool-grounded evidence E_t. It then updates the running profile ρ_t from the history h_t. The profile updater, the probe generator $\mathcal{G}$, and the scoring module σ are all realized by the model to defend f_θ, but with different prompt templates. In particular, the profile updater uses $\mathcal{P}_{\textsc{Profile}}$ to maintain ρ_t; the probe generator $\mathcal{G}$ uses $\mathcal{P}_{\textsc{GenerateProbe}}$ to produce a verification question q_t; and the scoring module σ uses $\mathcal{P}_{\textsc{Score}}$ to estimate sensitivity s_t, ambiguity Δ_t, and risk r_t. Based on these scores, AIV selects among the three actions in Eq. 10: refuse, verify, or assist. A verification budget B caps how many probes can be chained at one bottleneck before AIV must commit to either refusal or assistance. h_t, ρ_t, and the remaining budget are carried forward as state. In practice, we use thresholds $\tau_s = 0.5$ and $\tau_\Delta = 0.5$. We set the risk threshold to 0.8 and use a maximum of $B = 3$ verification questions when probing is triggered.

Algorithm 1 Active Intention Verification (AIV) with verify budget B

Require: History $h \leftarrow \emptyset$; profile $\rho \leftarrow \emptyset$; pending probe $q^\star \leftarrow \emptyset$; Verification budget $b \leftarrow B$.
Require: Defender model f_θ with prompt templates
$\mathcal{P}_{\textsc{Profile}},\ \mathcal{P}_{\textsc{Score}},\ \mathcal{P}_{\textsc{GenerateProbe}},\ \mathcal{P}_{\textsc{Assist}}$.
Require: Retrieval tool $\mathcal{R}$; thresholds $\tau_s, \tau_\Delta, \tau_r \in [0,1]$; budget $B \in \mathbb{N}$.

for each turn $t = 1, 2, \ldots$ **do**
  Receive a single user message u_t; $h \leftarrow h \cup \{u_t\}$.
  if $q^\star = \emptyset$ **then** ▷ Last interaction does not involve probe, hence we refresh the budget
    $b \leftarrow B$. ▷ We can ask at most B questions for each bottleneck
  end if
  $E_t \leftarrow \mathcal{R}(h, \rho)$. ▷ retrieve E_t
  $(\rho, s_t, \Delta_t, r_t) \leftarrow f_\theta\big(\mathcal{P}_{\textsc{Profile}}, \mathcal{P}_{\textsc{Score}};\ h, \rho, q^\star, E_t\big)$. ▷ Update Profile and Scoring
  $q^\star \leftarrow \emptyset$. ▷ Reset the probe question
  if $r_t \geq \tau_r$ **then** ▷ $d_t = 1$: refuse
    Output refusal; **continue**.
  end if
  if $b > 0$ **and** $s_t \geq \tau_s$ **and** $\Delta_t \geq \tau_\Delta$ **then** ▷ $d_t = \mathsf{V}$: verify
    $q_t \leftarrow f_\theta\big(\mathcal{P}_{\textsc{GenerateProbe}};\ h, \rho, E_t\big)$,
    with $q_t \in \mathcal{Q}_{\text{clar}} \cup \mathcal{Q}_{\text{cred}} \cup \mathcal{Q}_{\text{trap}}$.
    Output q_t; $h \leftarrow h \cup \{q_t\}$.
    $q^\star \leftarrow q_t$; $b \leftarrow b - 1$; **continue**.
  end if
  Output $a_t \leftarrow f_\theta\big(\mathcal{P}_{\textsc{Assist}};\ h, \rho, E_t\big)$. ▷ $d_t = 0$: assist
end for

5.3 Theoretical Guarantee

We formalize when one verification step overcomes the evidence bottleneck and improves over passive defenses. Section 5.2 motivates treating the probe response u_{t+1} as additional evidence separating benign from unsafe intent, which we capture with the following assumption.

Assumption 1 (Probe-induced safe partition)**.** *There exists $\eta \in [0, 1)$ such that, for every $h_t \in \mathcal{B}_\epsilon$:*

(i) (Safety of $\mathcal{A}_t$) *For all $z \in \mathcal{Z}$ with $\phi(z) = 1$,*

$$\mathbb{P}\big(\, S_{t+1}(z) = 1 \,\big|\, u_{t+1} \in \mathcal{A}_t,\ h_t,\ z \,\big) \ \leq\ \eta. \tag{11}$$

(ii) (Coverage of $\mathcal{A}_t$)

$$\mathbb{P}\big(\, u_{t+1} \in \mathcal{A}_t \,\big|\, y = 0,\ h_t \,\big) \ \geq\ 1 - \eta. \tag{12}$$

Assumption 1 encodes two desired effects of probing. First, *coverage*: benign users typically respond in a way that lands in $\mathcal{A}_t$ (part (ii)). Second, *safety*: even conditional on landing in $\mathcal{A}_t$, an unsafe user still receives unsafe assistance with probability at most η (part (i)). Under this assumption, we can bound AIV's expected loss after verification and compare it to the passive bottleneck lower bound. We provide empirical evidence of Assumption 1 in Appendix C.3.

Proposition 3. *Under Assumption 1, AIV's loss at the first post-verification decision turn satisfies*

$$\mathbb{E}[\, L_{t+1} \cdot \mathbf{1}\{h_t \in \mathcal{B}_\epsilon\}\,] \ \leq\ \eta \cdot (\alpha\, m_0 + \beta\, m_1), \tag{13}$$

where L_{t+1} is given by Equation 3. In contrast, Proposition 2 gives

$$\mathbb{E}[\, L_t \cdot \mathbf{1}\{h_t \in \mathcal{B}_\epsilon\}\,] \ \geq\ e^{-2\epsilon} \cdot \min(\alpha\, m_0,\ \beta\, m_1). \tag{14}$$

Hence, whenever

$$\eta \ <\ \frac{e^{-2\epsilon} \cdot \min(\alpha\, m_0,\ \beta\, m_1)}{\alpha\, m_0 + \beta\, m_1}, \tag{15}$$

AIV strictly improves the lower bound of total cost.

Remark 1 (Why L_{t+1}, not L_t)**.** *The loss L_t is incurred only when the defender commits to* assist *($d_t = 0$) or* refuse *($d_t = 1$). At a bottleneck, AIV instead chooses* verify *($d_t = \mathsf{V}$), so $L_t = 0$ at that turn. AIV's first non-zero loss occurs at turn $t + 1$, after observing (q_t, u_{t+1}). Accordingly, we compare passive's $\mathbb{E}[L_t\, \mathbf{1}\{h_t \in \mathcal{B}_\epsilon\}]$ to AIV's $\mathbb{E}[L_{t+1}\, \mathbf{1}\{h_t \in \mathcal{B}_\epsilon\}]$: each policy's loss at its first decision turn after reaching a bottleneck.*

6 EXPERIMENTS

6.1 EXPERIMENT SETUP

Datasets. Following prior work, we evaluate on 500 tasks from AdvBench (Zou et al., 2023). AdvBench consists of harmful user queries spanning physical harm, cyber abuse, fraud, misinformation, illegal activity, coercion, and discriminatory harassment. To measure both robustness *and* benign helpfulness, we create a paired benign set, which we call BenignPairBench, by constructing real-world, safety-oriented questions from queries in AdvBench. For example, "Give step-by-step instructions for how to steal someone's identity" becomes "Give step-by-step instructions for how to protect someone from identity theft." We then implement a benign user persona that is explicitly instructed to ask for defense-oriented information in multi-turn conversations, and mimic benign users in exhibiting ambiguity and under-specification. Additionally, we run further experiments to test the generalizability and robustness of our method on JailbreakBench (Chao et al., 2024) and ORBench (Cui et al., 2024). Details are summarized in Appendix C.2.

Models. We evaluate open-source models (Qwen2.5-7B-Instruct (Yang et al., 2024a), Llama-3-8B-Instruct (Grattafiori et al., 2024), Mistral-7B-Instruct-v0.2 (Jiang et al., 2023)) and closed-source models (GPT-4o, GPT-5.4, Claude-Sonnet-4.6). Due to API cost, we run 100 tasks for GPT-5.4 and Claude-Sonnet-4.6.

Attacks. We test multi-turn attacks including Crescendo (Russinovich et al., 2025), intention deception (Wang et al., 2026), Opposite-Day (Li et al., 2024), and Chain-of-Attack (Yang et al., 2024b). Consistent with (Wang et al., 2026) and our replications, intention deception is the only one that reliably jailbreaks frontier models (GPT-5.4, Claude-Sonnet-4.6); we therefore report frontier-model results only for this attack.

Defenses. We compare against X-Boundary (Lu et al., 2025), Neural Barrier Functions (Hu et al., 2025), and Temporal Context Awareness (Kulkarni and Namer, 2025). Since X-Boundary requires white-box access to the defended model, we evaluate it only on open-source models.

Metrics. We report Defense Success Rate (DSR) and Over-refusal Rate (OR). For DSR, we follow StrongREJECT (Souly et al., 2024) and use a rubric-based LLM judge that marks an attack successful only if the response is both compliant and sufficiently specific/convincing. For OR, we run BenignPairBench and measure the refusal rate. We use gpt-4o as the judge model.

6.2 EXPERIMENT RESULTS

Tables 1 and 2 summarizes the results on AdvBench and BenignPairBench. Table 1 reports defense success rate (DSR), the fraction of multi-turn attack attempts that are prevented, and Table 2 reports over-refusal rate (OR), the fraction of benign BenignPairBench queries that are incorrectly refused.

Our method effectively prevents jailbreaks. Our method is the top performer in every evaluated model–attack pair. Across the 18 reported rows, DSR remains high, from 0.826 up to 0.980, while passive baselines are brittle and can fail catastrophically on attacks that conceal intent across turns. The clearest example is *Intention Deception*: on open-source models, TCA, NBF, and X-Boundary top out at 0.216, 0.156, and 0.172 on Qwen2.5-7B and drop further on Mistral-7B and Llama-3-8B, whereas our method reaches 0.898, 0.854, and 0.826. This supports the central hypothesis of AIV: when each individual turn is designed to look benign, the defender must actively gather disambiguating evidence rather than scoring the attacker-chosen trajectory. The same pattern holds on closed-source models under Intention Deception, with DSR of 0.922 on GPT-4o, 0.92 on GPT-5.4, and 0.98 on Claude Sonnet 4.6.

Our method avoids refusing benign queries. Our method has the lowest OR on every model, from 0.022 to 0.05, meaning benign users are rarely blocked. In contrast, NBF refuses a large fraction of benign queries across models, which suggests that part of its apparent safety comes from broad suppression instead of intent-aware intervention.

Ablations. The *Ours-no-Q* variant provides strong evidence that active verification is essential as removing probe questions reduces our method to a passive defense, and DSR collapses under Intention Deception, for example from 0.898 to 0.152 on Qwen2.5-7B, from 0.854 to 0.118 on Llama-3-8B, and from 0.826 to 0.128 on Mistral-7B, while also increasing OR on every model.

Model	Attack	TCA	NBF	X-Boundary	Ours	Ours-no-Q	Ours-no-RAG	Ours-no-profile
Qwen2.5-7B	Crescendo	0.490	0.712	0.766	**0.972**	0.506	0.786	0.364
Qwen2.5-7B	Intention Deception	0.216	0.156	0.172	**0.898**	0.152	0.734	0.452
Qwen2.5-7B	Opposite-Day	0.478	0.866	0.824	**0.980**	0.498	0.884	0.780
Qwen2.5-7B	Chain-of-Attack	0.426	0.652	0.604	**0.966**	0.478	0.870	0.402
Llama-3-8B	Crescendo	0.450	0.696	0.804	**0.948**	0.492	0.840	0.662
Llama-3-8B	Intention Deception	0.192	0.144	0.136	**0.854**	0.118	0.726	0.432
Llama-3-8B	Opposite-Day	0.514	0.862	0.830	**0.976**	0.784	0.928	0.808
Llama-3-8B	Chain-of-Attack	0.672	0.746	0.588	**0.970**	0.658	0.878	0.734
Mistral-7B	Crescendo	0.472	0.670	0.758	**0.914**	0.452	0.786	0.516
Mistral-7B	Intention Deception	0.134	0.152	0.120	**0.826**	0.128	0.772	0.348
Mistral-7B	Opposite-Day	0.536	0.810	0.794	**0.890**	0.548	0.828	0.624
Mistral-7B	Chain-of-Attack	0.626	0.702	0.496	**0.852**	0.612	0.790	0.684
GPT-4o	Crescendo	0.536	0.648	–	**0.940**	0.578	0.812	0.524
GPT-4o	Intention Deception	0.214	0.192	–	**0.922**	0.220	0.768	0.596
GPT-4o	Opposite-Day	0.612	0.836	–	**0.954**	0.594	0.824	0.708
GPT-4o	Chain-of-Attack	0.630	0.604	–	**0.938**	0.624	0.806	0.702
GPT-5.4	Intention Deception	0.54	0.67	–	**0.92**	0.60	0.85	0.79
Claude Sonnet 4.6	Intention Deception	0.65	0.72	–	**0.98**	0.63	0.94	0.76

Table 1: Defense Success Rate (DSR) on multi-turn jailbreak attacks across models (higher is better). Ours-no-Q indicates the ablation where we drop Question Generation(Q). For each row, the best-performing defense method is marked in bold. For GPT-5.4 and Claude Sonnet 4.6 we evaluate 100 tasks due to API cost. We omit non-effective frontier-model attacks as described in Section 6.

Model	TCA	NBF	X-Boundary	Ours	Ours-no-Q	Ours-no-RAG	Ours-no-profile
Qwen2.5-7B	0.102	0.614	0.090	**0.022**	0.138	0.040	0.086
Llama-3-8B	0.158	0.568	0.124	**0.030**	0.146	0.052	0.094
Mistral-7B	0.096	0.484	0.118	**0.042**	0.110	0.068	0.104
GPT-4o	0.136	0.524	–	**0.038**	0.108	0.054	0.072
GPT-5.4	0.13	0.54	–	**0.05**	0.14	0.08	0.08
Claude Sonnet 4.6	0.12	0.47	–	**0.03**	0.09	0.06	0.11

Table 2: Over-refusal rate (OR) on BenignPairBench (lower is better). Ours-no-Q indicates the ablation where we drop Question Generation(Q). For each row, the best-performing defense method is marked in bold. For GPT-5.4 and Claude Sonnet 4.6 we evaluate 100 tasks due to API cost.

Removing profile memory (*Ours-no-profile*) reduces DSR more moderately but consistently across attacks. This suggests that the probe questions themselves already help shape the trajectory and elicit intent-revealing cues, but maintaining a stable cross-turn profile is important for aggregating these cues and recognizing slow intent drift. Removing retrieval (*Ours-no-RAG*) also degrades DSR, indicating that parameterized knowledge alone is insufficient for reliable intent verification in dual-use domains, and tool-grounded evidence can add complementary signals beyond what is stored in model parameters, improving both disambiguation and calibration once probing is triggered.

7 Conclusion

Multi-turn jailbreak defense is fundamentally an intent inference problem: attackers can spread and selectively reveal evidence across turns so each request appears benign, leaving purely passive safeguards with an irreducible error floor. We formalized this *evidence bottleneck* and reframed multi-turn defense as a sequential decision problem under asymmetric information. We then proposed *Active Intention Verification* (AIV), the first interactive defense that improves defense success while reducing over-refusals on a matched benign benchmark; ablations indicate that active questioning drives robustness, with profile memory and retrieval providing complementary gains. More broadly, our results suggest that the key limitation in multi-turn safety is not turn-level classification, but making hidden intent identifiable early enough. This points toward verifiable dialogue protocols, interactions that actively elicit, constrain, and ground evidence, rather than static filters applied to whatever the attacker chooses to reveal.

AI USE STATEMENT

In this work, we used generative AI tools to generate synthetic datasets, assist with the implementation of methods, and assist with translation.

We did not use generative AI tools to develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs, propose or refine hypotheses, design or provide feedback on research methodology or experiments, clean or reformat datasets, support qualitative or thematic data analysis, or interpret results, where applicable.

Additionally, we used generative AI tools to source and search for information and to edit the research paper for clarity, readability, and language quality.

All AI-assisted outputs were reviewed and verified by the authors. Synthetic data generated with AI assistance were checked for consistency with the intended data structure and study requirements. AI-assisted implementation was reviewed and tested for correctness before use. Translations and language edits were manually reviewed against the intended meaning and revised where necessary. Information identified through AI-assisted searching was verified against appropriate original or scholarly sources before being incorporated into the work. We take full responsibility for the final content of this work, including all text, claims, data, code, and other artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work studies defenses against multi-turn jailbreak attacks and therefore necessarily involves safety-sensitive prompts and adversarial interactions. These materials are used exclusively to evaluate model robustness and defensive behavior. Our experiments use established jailbreak and safety benchmarks together with a matched benign benchmark to measure both defense success and over-refusal. The proposed Active Intention Verification framework is designed to reduce unsafe assistance by actively disambiguating user intent, but it may still produce both false positives and false negatives and should not be interpreted as providing a complete safety guarantee.

Research on jailbreak defenses is inherently dual-use: descriptions of attack behaviors and defensive failure modes may also inform adversaries. We therefore focus our presentation on defensive principles, aggregate evaluation, and the evidence needed for scientific reproducibility, while avoiding unnecessary disclosure of operational details that could materially facilitate misuse. Generative AI was also used in parts of the research workflow as described separately in our AI Use Statement. The authors take responsibility for the experimental design, analysis, conclusions, and release decisions in this work.

REFERENCES

Justin Albrethsen, Yash Datta, Kunal Kumar, and Sharath Rajasekar. Deepcontext: Stateful real-time detection of multi-turn adversarial intent drift in llms. *arXiv preprint arXiv:2602.16935*, 2026.

Patrick Chao, Edoardo Debenedetti, Alexander Robey, Maksym Andriushchenko, Francesco Croce, Vikash Sehwag, Edgar Dobriban, Nicolas Flammarion, George J Pappas, Florian Tramer, et al. Jailbreakbench: An open robustness benchmark for jailbreaking large language models. *Advances in Neural Information Processing Systems*, 37:55005–55029, 2024.

Justin Cui, Wei-Lin Chiang, Ion Stoica, and Cho-Jui Hsieh. Or-bench: An over-refusal benchmark for large language models. *arXiv preprint arXiv:2405.20947*, 2024.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Hanjiang Hu, Alexander Robey, and Changliu Liu. Steering dialogue dynamics for robustness against multi-turn jailbreaking attacks. *arXiv preprint arXiv:2503.00187*, 2025.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b. *CoRR*, abs/2310.06825, 2023. doi: 10.48550/ARXIV.2310.06825. URL https://doi.org/10.48550/arXiv.2310.06825.

Prashant Kulkarni and Assaf Namer. Temporal context awareness: A defense framework against multi-turn manipulation attacks on large language models. In *2025 IEEE Conference on Artificial Intelligence (CAI)*, pages 930–935. IEEE, 2025.

Nathaniel Li, Ziwen Han, Ian Steneker, Willow Primack, Riley Goodside, Hugh Zhang, Zifan Wang, Cristina Menghini, and Summer Yue. Llm defenses are not robust to multi-turn human jailbreaks yet. *arXiv preprint arXiv:2408.15221*, 2024.

Xiaoya Lu, Dongrui Liu, Yi Yu, Luxin Xu, and Jing Shao. X-boundary: Establishing exact safety boundary to shield llms from multi-turn jailbreaks without compromising usability. *arXiv preprint arXiv:2502.09990*, 2025.

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. *arXiv preprint arXiv:2504.13203*, 2025.

Mark Russinovich, Ahmed Salem, and Ronen Eldan. Great, now write an article about that: The crescendo {Multi-Turn}{LLM} jailbreak attack. In *34th USENIX Security Symposium (USENIX Security 25)*, pages 2421–2440, 2025.

Xinyue Shen, Zeyuan Chen, Michael Backes, Yun Shen, and Yang Zhang. " do anything now": Characterizing and evaluating in-the-wild jailbreak prompts on large language models. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, pages 1671–1685, 2024.

Alexandra Souly, Qingyuan Lu, Dillon Bowen, Tu Trinh, Elvis Hsieh, Sana Pandey, Pieter Abbeel, Justin Svegliato, Scott Emmons, Olivia Watkins, et al. A strongreject for empty jailbreaks. *Advances in Neural Information Processing Systems*, 37:125416–125440, 2024.

Xiongtao Sun, Deyue Zhang, Dongdong Yang, Quanchen Zou, and Hui Li. Multi-turn context jailbreak attack on large language models from first principles. *arXiv preprint arXiv:2408.04686*, 2024.

Xinhe Wang, Katia Sycara, and Yaqi Xie. Jailbreaking frontier foundation models through intention deception. *arXiv preprint arXiv:2604.24082*, 2026.

Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. Jailbroken: How does llm safety training fail? *Advances in neural information processing systems*, 36:80079–80110, 2023.

Zixuan Weng, Xiaolong Jin, Jinyuan Jia, and Xiangyu Zhang. Foot-in-the-door: A multi-turn jailbreak for llms. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 1939–1950, 2025.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024a.

Xikang Yang, Xuehai Tang, Songlin Hu, and Jizhong Han. Chain of attack: a semantic-driven contextual multi-turn attacker for llm. *arXiv preprint arXiv:2405.05610*, 2024b.

Sibo Yi, Yule Liu, Zhen Sun, Tianshuo Cong, Xinlei He, Jiaxing Song, Ke Xu, and Qi Li. Jailbreak attacks and defenses against large language models: A survey. *arXiv preprint arXiv:2407.04295*, 2024.

Andy Zhou and Ron Arel. Tempest: Autonomous multi-turn jailbreaking of large language models with tree search. *arXiv preprint arXiv:2503.10619*, 2025.

Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*, 2023.

A PROOFS

A.1 PROOF OF PROPOSITION 1

Proof. Fix a deterministic rule d and let $A := \{h \in B_\epsilon : d(h) = 1\}$. Write $\mu_y(\cdot) := P(h_t \in \cdot \mid y)$ for the intent-conditional law of the history at turn t.

By the first clause of Definition 1, every $h \in B_\epsilon$ satisfies $|\Lambda(h)| \le \epsilon$, i.e.,

$$e^{-\epsilon}\, P(h \mid y = 0) \;\le\; P(h \mid y = 1) \;\le\; e^{\epsilon}\, P(h \mid y = 0).$$

Integrating these pointwise inequalities over any measurable $E \subseteq B_\epsilon$ gives

$$e^{-\epsilon}\, \mu_0(E) \;\le\; \mu_1(E) \;\le\; e^{\epsilon}\, \mu_0(E). \tag{16}$$

In particular, $\mu_1(B_\epsilon) \le e^{\epsilon}\, \mu_0(B_\epsilon)$ and $\mu_1(B_\epsilon \setminus A) \ge e^{-\epsilon}\, \mu_0(B_\epsilon \setminus A)$.

The false-negative rate on B_ϵ is $P(d(h_t) = 0 \mid y = 1,\, h_t \in B_\epsilon) = \mu_1(B_\epsilon \setminus A)/\mu_1(B_\epsilon)$. Combining the two inequalities from Step 1,

$$\frac{\mu_1(B_\epsilon \setminus A)}{\mu_1(B_\epsilon)} \;\ge\; \frac{e^{-\epsilon}\, \mu_0(B_\epsilon \setminus A)}{e^{\epsilon}\, \mu_0(B_\epsilon)} \;=\; e^{-2\epsilon}\left(1 - \frac{\mu_0(A)}{\mu_0(B_\epsilon)}\right).$$

The false-positive rate on B_ϵ is $P(d(h_t) = 1 \mid y = 0,\, h_t \in B_\epsilon) = \mu_0(A)/\mu_0(B_\epsilon)$. Adding it to the inequality from Step 2 and letting $\alpha_0 := \mu_0(A)/\mu_0(B_\epsilon) \in [0, 1]$,

$$\frac{\mu_0(A)}{\mu_0(B_\epsilon)} + \frac{\mu_1(B_\epsilon \setminus A)}{\mu_1(B_\epsilon)} \;\ge\; \alpha_0 + e^{-2\epsilon}(1 - \alpha_0) \;=\; e^{-2\epsilon} + (1 - e^{-2\epsilon})\alpha_0 \;\ge\; e^{-2\epsilon},$$

since $1 - e^{-2\epsilon} \ge 0$ and $\alpha_0 \ge 0$. This is the claimed bound for deterministic d.

A randomized rule d is a measurable map $h \mapsto P(d(h) = 1 \mid h)$ with values in $[0, 1]$. Both conditional error rates are linear functionals of this map, so their sum is linear in d. The set of randomized rules is the convex hull of the deterministic rules, hence the linear functional attains its minimum at an extreme point (a deterministic rule), where Steps 1–3 apply. The bound therefore persists. □

A.2 PROOF OF PROPOSITION 2

Proof. Let

$$a := \Pr(d_t = 1 \,|\, y = 0,\, h_t \in B_\epsilon)\,, \qquad b := \Pr(d_t = 0 \,|\, y = 1,\, h_t \in B_\epsilon)\,. \tag{17}$$

By Proposition 1,

$$a + b \ge e^{-2\epsilon}. \tag{18}$$

From Equation 3, the expected loss on the bottleneck set is

$$\begin{aligned}\mathbb{E}[L_t \mathbf{1}\{h_t \in B_\epsilon\}] = \alpha \Pr(y = 0, d_t = 1, h_t \in B_\epsilon) \\ + \beta \mathbb{E}[\mathbf{1}\{y = 1, d_t = 0, h_t \in B_\epsilon\} S_t(z)]\,.\end{aligned} \tag{19}$$

The false-positive term satisfies

$$\Pr(y = 0, d_t = 1, h_t \in B_\epsilon) = m_0 a. \tag{20}$$

For the false-negative term, define

$$E := \{y = 1, d_t = 0, h_t \in B_\epsilon\}. \tag{21}$$

For every realized $h_t \in B_\epsilon$ on the event E, $d_t = 0$ implies $\exists s \ge t : d_s = 0$. By Definition 1,

$$\Pr(S_t(z) = 1 \,|\, y = 1, h_t, d_t = 0) = 1. \tag{22}$$

Averaging over h_t conditional on E therefore gives

$$\Pr(S_t(z) = 1 \,|\, E) = 1. \tag{23}$$

Hence,

$$\mathbb{E}[\mathbf{1}_E S_t(z)] = \Pr(E) = m_1 b. \tag{24}$$

Substituting Equations 20 and 24 into Equation 19 yields

$$\mathbb{E}[L_t \mathbf{1}\{h_t \in B_\epsilon\}] = \alpha m_0 a + \beta m_1 b. \tag{25}$$

Since $a, b \geq 0$,

$$\alpha m_0 a + \beta m_1 b \geq \min(\alpha m_0, \beta m_1)(a + b) \tag{26}$$

$$\geq e^{-2\epsilon} \min(\alpha m_0, \beta m_1), \tag{27}$$

where the second inequality follows from Equation 18. This proves the claim. □

A.3 Proof of Proposition 3

Proof. We consider the case that only one question is asked for the bottleneck and hence the first non-verifying decision occurs at $t+1$ with history $h_{t+1} = (h_t, q_t, u_{t+1})$. The case when multiple questions are asked simply follows by recursion. We treat the case $d(h_{t+1}) \in \{0, 1\}$. By construction of $(\mathcal{A}_t, \mathcal{F}_t)$,

$$d_{t+1} = 0 \iff u_{t+1} \in \mathcal{A}_t, \qquad d_{t+1} = 1 \iff u_{t+1} \in \mathcal{F}_t.$$

Substituting into Equation 3 at turn $t+1$:

$$L_{t+1} = \alpha\, \mathbf{1}[y{=}0]\, \mathbf{1}[u_{t+1} \in \mathcal{F}_t] + \beta\, \mathbf{1}[y{=}1]\, \mathbf{1}[u_{t+1} \in \mathcal{A}_t]\, S_{t+1}(z).$$

Taking expectation and conditioning on $\{h_t \in \mathcal{B}_\epsilon\}$,

$$\mathbb{E}[L_{t+1} \cdot \mathbf{1}\{h_t \in \mathcal{B}_\epsilon\}] = \alpha\, m_0 \cdot P_{\mathrm{FP}} + \beta\, m_1 \cdot P_{\mathrm{FN}},$$

where

$$P_{\mathrm{FP}} := \mathbb{P}(u_{t+1} \in \mathcal{F}_t \mid y{=}0,\, h_t \in \mathcal{B}_\epsilon), \quad P_{\mathrm{FN}} := \mathbb{E}\big[\mathbf{1}[u_{t+1} \in \mathcal{A}_t]\, S_{t+1}(z) \,\big|\, y{=}1,\, h_t \in \mathcal{B}_\epsilon\big].$$

Now we bound P_{FP}. $\mathcal{A}_t$ and $\mathcal{F}_t$ are disjoint, so $\mathcal{F}_t \subseteq \mathcal{U} \setminus \mathcal{A}_t$:

$$P_{\mathrm{FP}} \leq \mathbb{P}(u_{t+1} \notin \mathcal{A}_t \mid y{=}0,\, h_t \in \mathcal{B}_\epsilon) = 1 - \mathbb{P}(u_{t+1} \in \mathcal{A}_t \mid y{=}0,\, h_t \in \mathcal{B}_\epsilon) \leq \eta,$$

by Assumption 1(ii).

Now we bound P_{FN}. Condition on z via the tower rule:

$$P_{\mathrm{FN}} = \mathbb{E}_{z \mid y=1,\, h_t \in \mathcal{B}_\epsilon}[\mathbb{E}[\mathbf{1}[u_{t+1} \in \mathcal{A}_t]\, S_{t+1}(z) \mid z,\, y{=}1,\, h_t \in \mathcal{B}_\epsilon]].$$

Since $y = \phi(z)$, conditioning on z makes the event $\{y = 1\}$ redundant; the inner expectation factors as

$$\mathbb{P}(u_{t+1} \in \mathcal{A}_t \mid z,\, h_t \in \mathcal{B}_\epsilon) \cdot \mathbb{P}(S_{t+1}(z) = 1 \mid u_{t+1} \in \mathcal{A}_t,\, z,\, h_t \in \mathcal{B}_\epsilon) \leq 1 \cdot \eta = \eta,$$

where the last bound applies Assumption 1(i) (which holds because $\phi(z) = 1$ on the support of $\{y = 1\}$). Hence $P_{\mathrm{FN}} \leq \eta$.

Combining the above two,

$$\mathbb{E}[L_{t+1} \cdot \mathbf{1}\{h_t \in \mathcal{B}_\epsilon\}] \leq \alpha\, m_0\, \eta + \beta\, m_1\, \eta = \eta\, (\alpha\, m_0 + \beta\, m_1).$$

The strict-improvement condition follows by direct comparison with Proposition 2. □

B Additional Implementation Details

Profiling and Scoring Profile updating and scoring are performed jointly in a single structured call to the defended model. The call takes as input the retrieved evidence, the current user message, the user role recorded in previous turns, the scores from the previous turn, and the four most recent messages. It outputs the updated user and topic information together with the sensitivity, ambiguity, and risk scores, plausible legitimate and harmful interpretations of the request, and any red flags. The prompt provides explicit rubrics for each score. Sensitivity captures the inherent harmful potential

Model	Method	Additional Calls	Median E2E	P95 E2E	Total System Time
		Adversarial Cases			
Qwen2.5-7B	AIV	1.0	+13.53%	−9.18%	+11.18%
Qwen2.5-7B	TCA	0.9	+27.99%	+14.44%	+34.61%
Qwen2.5-7B	NBF	0.0	−99.72%	−39.87%	−71.99%
GPT-4o	AIV	1.0	−17.67%	+6.89%	−7.96%
GPT-4o	TCA	0.9	−23.69%	+12.73%	−5.41%
GPT-4o	NBF	0.0	−98.98%	−30.06%	−71.47%
		Benign Cases			
Qwen2.5-7B	AIV	1.0	+11.52%	+5.00%	+6.17%
Qwen2.5-7B	TCA	0.9	+24.82%	+11.31%	+16.70%
Qwen2.5-7B	NBF	0.0	−88.93%	−14.37%	−74.21%
GPT-4o	AIV	1.0	+10.25%	+12.04%	+10.15%
GPT-4o	TCA	0.9	+38.65%	+31.03%	+33.38%
GPT-4o	NBF	0.0	−3.84%	−12.41%	−24.74%

Table 3: Computational overhead relative to the undefended model (RAW) on adversarial and benign conversations. Each setting contains 100 tasks, with each task evaluated over a fixed 10-turn conversation. Additional calls are averaged per processed user turn. E2E latency and total system time are reported as percentage changes relative to RAW.

of the topic, independent of the user's claimed role or intent. Ambiguity captures how difficult it is to distinguish a legitimate interpretation from a malicious one given the available context. Risk captures the likelihood that a complete response would materially enable harm. Because profiling and scoring share one call, AIV adds approximately one model call per processed turn.

Retrieval. At each turn, AIV queries Wikipedia using the topic extracted in the previous profile update. It retrieves up to five articles, splits them into chunks, and encodes the chunks and the current message with a frozen all-MiniLM-L6-v2 encoder. The five chunks most similar to the current message are retained, provided their cosine similarity exceeds 0.25. These chunks are passed as evidence to the joint profile-and-scoring call.

Long-range context. Information from earlier turns is carried forward through the profile, which persists the user's recorded role, the topic, the previous scores, and the flagged inconsistencies across turns. The profile thus acts as a compressed memory of the full conversation. This lets AIV detect slow intent drift, where individual recent turns look benign but diverge from the role or objective established earlier, without re-reading the entire history at every turn. The Ours-no-profile ablation (Table 1) supports this design: removing the profile consistently reduces DSR across attacks, even though the four-message window is retained.

C ADDITIONAL EXPERIMENTS

C.1 COMPUTATIONAL OVERHEAD

We evaluate the practical computational overhead of different defense methods under both benign and adversarial interactions. Specifically, we sample 100 benign and 100 adversarial tasks and run each task as a fixed 10-turn conversation using Qwen2.5-7B-Instruct and GPT-4o. We compare AIV, TCA, and NBF against the undefended model (RAW). All methods process all 10 turns for each task, including turns following a refusal.

We report four efficiency metrics. *Additional calls* denotes the average number of model calls added per processed user turn relative to RAW. *Median E2E* and *P95 E2E* measure per-turn end-to-end latency, including retrieval and all model calls. *Total system time* is the cumulative processing time over the full evaluation, excluding user waiting time. All latency and system-time values are reported relative to RAW under the same model and interaction setting.

Model	Attack	TCA	NBF	X-Boundary	AIV
Qwen2.5-7B	Intention Deception	0.26	0.18	0.19	**0.95**
GPT-4o	Intention Deception	0.22	0.19	N/A	**0.93**
Llama3-8B	Intention Deception	0.17	0.22	0.15	**0.90**
Qwen2.5-7B	Crescendo	0.52	0.70	0.73	**0.93**
GPT-4o	Crescendo	0.57	0.61	N/A	**0.97**
Llama3-8B	Crescendo	0.43	0.65	0.76	**0.94**

Table 4: Defense success rate (DSR) on JailbreakBench under two multi-turn attacks. Higher is better.

Model	TCA	NBF	X-Boundary	AIV
Qwen2.5-7B	0.19	0.69	0.22	**0.06**
GPT-4o	0.14	0.57	N/A	**0.02**
Llama3-8B	0.16	0.63	0.13	**0.06**

Table 5: Refusal rate on the public OR-Bench. Lower is better. Each original benign query is expanded into a 10-turn topic-consistent conversation.

As shown in Table 3, AIV introduces one additional model call per processed turn, but its wall-clock overhead remains modest. On adversarial tasks, its total system time ranges from -7.96% to $+11.18\%$ relative to RAW; on benign tasks, the increase is limited to $+6.17\%$–$+10.15\%$. Its median per-turn latency is similarly bounded, ranging from -17.67% to $+13.53\%$ on adversarial inputs and approximately $+10\%$–$+12\%$ on benign inputs.

AIV also incurs lower practical overhead than TCA in the benign setting. AIV increases total system time by only 6.17%–10.15%, whereas TCA increases it by 16.70%–33.38%. This trend is also reflected in median E2E latency, where AIV incurs a 10.25%–11.52% increase compared with 24.82%–38.65% for TCA.

NBF exhibits substantial apparent speedups because it can terminate an interaction with a rejection without invoking the main generation model. Consequently, its reduced latency and total system time should not be interpreted as a directly comparable efficiency gain: these reductions are closely associated with its tendency to reject benign requests, as discussed in our benign-helpfulness evaluation.

C.2 Generalizability Across Benchmarks and Models

We further evaluate whether AIV generalizes beyond the AdvBench-based evaluation used in the main experiments. Specifically, we conduct additional experiments on JailbreakBench which tests defense against adversarial queries and OR-Bench which tests over-refusal. No dataset-specific tuning is applied.

For JailbreakBench, we evaluate the full set of 100 harmful behaviors under two multi-turn attacks: *Intention Deception* and *Crescendo*. For OR-Bench, which is originally a large-scale single-turn benign benchmark, we sample 100 benign queries spanning 10 categories. To adapt these queries to the multi-turn setting considered in our evaluation, we identify the underlying topic of each query and initialize a benign user agent that continues asking related questions about that topic. This produces a 10-turn benign conversation for each sampled query, rather than repeatedly presenting the original single-turn request.

We evaluate Qwen2.5-7B-Instruct, GPT-4o, and Llama3-8B-Instruct, and compare AIV against TCA, NBF, and X-Boundary. X-Boundary is not evaluated on GPT-4o because it requires white-box access to internal model representations and parameters.

As shown in Table 4, AIV maintains strong robustness on JailbreakBench across both attack strategies and all three evaluated models. Under Intention Deception, AIV achieves a DSR of 0.90–0.95,

Model	Benign Coverage $\widehat{P}(u_{t+1} \in A_t \mid y = 0) \uparrow$	Conditional ASR $\widehat{P}(S_{t+1} = 1 \mid u_{t+1} \in A_t, y = 1) \downarrow$	Empirical $\hat{\eta} \downarrow$
Qwen2.5-7B	0.986	0.041	0.041
GPT-4o	0.991	0.029	0.029
Llama3-8B	0.984	0.026	0.026
Mistral-7B	0.988	0.047	0.047

Table 6: Empirical evaluation of Assumption 1 on 500 AdvBench task topics. Benign coverage measures the fraction of benign probe responses that fall in the acceptance region A_t, while conditional ASR measures the fraction of malicious responses admitted into A_t that subsequently lead to unsafe assistance.

while under Crescendo it achieves 0.93–0.97. These results indicate that the effectiveness of AIV is not restricted to the attack patterns or benchmark used in the original evaluation.

The results also generalize across model families. Across Qwen2.5-7B-Instruct, GPT-4o, and Llama3-8B-Instruct, AIV consistently achieves higher DSR than TCA, NBF, and X-Boundary on both JailbreakBench attacks. In particular, the improvement remains substantial on Llama3-8B-Instruct, providing evidence that the defense does not depend on a single underlying model family.

Table 5 further shows that this robustness does not come at the cost of excessive refusal on benign inputs. On the public OR-Bench, AIV yields refusal rates of only 0.02–0.06 across the three models. In comparison, NBF refuses 0.57–0.69 of benign conversations, while TCA and X-Boundary exhibit refusal rates between 0.13 and 0.22. Together, these results suggest that AIV generalizes across benchmarks, attack strategies, and model families while preserving benign utility.

C.3 Empirical Validation of Prop.3 Assumption 1

We empirically evaluate the two operational requirements of Assumption 1 on all 500 AdvBench task topics. For the unsafe-intent condition, we use trajectories generated under the Intention Deception attack. For the benign-coverage condition, we use corresponding trajectories generated by a benign user persona on the same task topics.

The response partition in Assumption 1 is induced by AIV's post-probe decision rule. Specifically, we treat A_t as the set of probe responses after which AIV proceeds to assistance. From the observed verification events, we estimate the benign coverage as

$$\widehat{P}(u_{t+1} \in A_t \mid y = 0)\,, \tag{28}$$

and the conditional attack success rate among malicious responses that enter the same acceptance region as

$$\widehat{P}(S_{t+1} = 1 \mid u_{t+1} \in A_t,\ y = 1)\,. \tag{29}$$

Assumption 1 requires the same η to upper-bound the safety error while requiring a benign response to enter A_t with probability at least $1 - \eta$. We therefore report the aggregate empirical proxy

$$\hat{\eta} = \max\left\{1 - \widehat{P}(u_{t+1} \in A_t \mid y = 0)\,,\ \widehat{P}(S_{t+1} = 1 \mid u_{t+1} \in A_t,\ y = 1)\right\}. \tag{30}$$

As shown in Table 6, A_t covers 98.4%–99.1% of benign probe responses across the four evaluated models. At the same time, only 2.6%–4.7% of malicious probe responses that enter A_t subsequently result in unsafe assistance. The resulting empirical $\hat{\eta}$ ranges from 0.026 to 0.047.

These results provide empirical support for both operational conditions in Assumption 1. Benign probe responses enter the acceptance region with high probability, while malicious responses admitted into the same region have a low probability of leading to unsafe assistance. They also make explicit how the abstract partition in Proposition 3 is instantiated by AIV's post-probe decision rule in practice.

C.4 Reliability Analysis of AIV Components

We further examine the reliability of the three core components underlying AIV's decisions: the scoring mechanism, the profile-update mechanism, and active probing. These analyses complement

Target Model	Dimension	Agreement
GPT-4o	Sensitivity	0.923
GPT-4o	Ambiguity	0.935
GPT-4o	Risk	0.980
Qwen2.5-7B	Sensitivity	0.931
Qwen2.5-7B	Ambiguity	0.927
Qwen2.5-7B	Risk	0.970
Pooled	**Sensitivity**	**0.927**
Pooled	**Ambiguity**	**0.931**
Pooled	**Risk**	**0.975**

Table 7: Agreement of AIV's sensitivity, ambiguity, and risk decisions with independent annotations. Continuous AIV scores are converted into high/low decisions using the default thresholds employed by the defense.

the end-to-end robustness and benign-utility results by evaluating whether the intermediate signals used by AIV behave consistently with their intended semantics.

C.4.1 Scoring Reliability

We first evaluate the reliability of AIV's sensitivity, ambiguity, and risk scores. In addition to the one-at-a-time threshold sensitivity analysis, we directly examine whether the scores support reliable decisions at the default thresholds used by AIV.

We evaluate GPT-4o and Qwen2.5-7B-Instruct on 500 benign, 500 Intention Deception, and 500 Crescendo instances for each model, yielding 3,000 instances in total. Claude Sonnet 5 independently annotates each instance according to the same definitions of sensitivity, ambiguity, and risk used by AIV, while remaining blinded to AIV's predicted scores and selected actions.

For each dimension, we convert AIV's continuous score into a binary high/low decision using the corresponding default threshold: τ_s for sensitivity, τ_Δ for ambiguity, and τ_r for risk. Using the independent annotations as reference labels, we report the agreement

As shown in Table 7, all three scoring dimensions exhibit strong agreement with the independent annotations.

These results indicate that the thresholded sensitivity, ambiguity, and risk scores used to determine assistance, verification, and refusal are closely aligned with the independently annotated labels.

C.4.2 Profile-Update Reliability

We next evaluate whether the red flags and hypotheses produced by AIV's profile-update mechanism are consistent with the underlying hidden objective of an adversarial conversation. We analyze the complete 500-task Intention Deception and Crescendo evaluations for GPT-4o and Qwen2.5-7B-Instruct.

For each adversarial trajectory, Claude Sonnet 5 is given the finding produced by the profile-update mechanism together with the ground-truth hidden attack objective and independently judges whether the two are consistent. We report both the resulting consistency rate and AIV's defense success rate (DSR) conditioned on the profile finding being judged consistent with the hidden objective.

Table 8 shows that 85.8%–93.6% of profile findings are judged consistent with the underlying hidden attack objective across the two models and two attack strategies. This indicates that the profile-update mechanism frequently captures the specific latent risk developing across the conversation.

Consistent profile findings are also strongly associated with successful defense. Conditioned on the profile finding being consistent with the hidden objective, AIV achieves a DSR of 99.5%–100%. This result supports the role of the profile as an informative intermediate representation for subsequent verification and safety decisions.

Target Model	Attack	Objective Consistency	DSR | Consistent Finding
GPT-4o	Intention Deception	0.872	0.998
GPT-4o	Crescendo	0.936	1.000
Qwen2.5-7B	Intention Deception	0.858	0.995
Qwen2.5-7B	Crescendo	0.924	1.000

Table 8: Reliability of AIV's profile-update mechanism on adversarial trajectories. Objective consistency denotes the fraction of profile findings judged consistent with the underlying hidden attack objective. The final column reports AIV's defense success rate conditioned on a consistent profile finding.

Target Model	Dataset	Before Probe	After Response	Reduction
GPT-4o	Benign	0.738	0.393	0.345
GPT-4o	Intention Deception	0.761	0.410	0.351
GPT-4o	Crescendo	0.712	0.355	0.357
Qwen2.5-7B	Benign	0.729	0.405	0.324
Qwen2.5-7B	Intention Deception	0.774	0.456	0.318
Qwen2.5-7B	Crescendo	0.737	0.399	0.338

Table 9: Change in AIV's ambiguity score before and after active verification. Each row aggregates all probes generated for the corresponding target model and evaluation setting. Reduction is the difference between the mean ambiguity score before the probe and after the user's response.

C.4.3 Probe Effectiveness

Finally, we examine whether active probes reduce unresolved ambiguity in the conversation. We collect all probes generated during the complete 500-task benign, Intention Deception, and Crescendo evaluations for GPT-4o and Qwen2.5-7B-Instruct. Using the ambiguity dimension evaluated above, we compare AIV's ambiguity score immediately before each probe with the score after the user responds to that probe.

As shown in Table 9, ambiguity decreases after the user's response in every evaluated model and interaction setting. The mean reduction ranges from 0.318 to 0.357. Aggregated across all six settings, the mean ambiguity score decreases from approximately 0.742 before probing to 0.403 after the response.

The reduction is consistent across both benign and adversarial conversations, indicating that active probing elicits additional information that resolves uncertainty about the user's operational context. The downstream behavior after probing is analyzed separately in Appendix C.3, where we measure benign coverage of the post-probe acceptance region and the conditional attack success rate after AIV proceeds to assistance.

C.5 Robustness to Protocol-Aware Adaptive Attacks

The multi-turn attacks used in our main evaluation are not explicitly designed with knowledge of AIV's verification protocol. We therefore conduct an additional evaluation against protocol-aware adaptive attackers that specifically anticipate and respond to AIV's probes.

We build these adaptive attacks on Intention Deception, which already provides a strong basis for protocol-aware adaptation. In particular, the attack constructs a consistent benign cover story and maintains credential and intent consistency across turns. We introduce two additional adaptations that explicitly target AIV's verification mechanism.

In *Intention Deception + (a)*, the attacker anticipates likely clarification and credential probes, prepares responses that are consistent with its benign cover story in advance, and retrieves these responses when the corresponding probes are encountered. This adaptation tests whether pre-staged, internally consistent answers can reduce the effectiveness of active clarification.

Model	ID + (a)	ID + (b)	ID + (a) + (b)	Original ID
Qwen2.5-7B	0.83	0.86	0.82	0.88
GPT-4o	0.87	0.91	0.85	0.93
Llama3-8B	0.82	0.85	0.78	0.85
Mistral-7B	0.79	0.81	0.78	0.82
Claude Sonnet 4.6	0.95	0.97	0.94	0.98
Average	**0.852**	**0.880**	**0.834**	**0.892**

Table 10: Defense success rate (DSR) of AIV against protocol-aware variants of Intention Deception on a 100-task AdvBench subset. Variant (a) prepares consistent answers to anticipated clarification and credential probes; variant (b) avoids or redirects trap probes; and (a)+(b) combines both adaptations. The original Intention Deception results are evaluated on the same 100 tasks. Higher is better.

In *Intention Deception + (b)*, the attacker is informed of the purpose of AIV's trap probes and is instructed to decline or redirect suggestions that would be useful for the hidden harmful objective but irrelevant to the stated benign goal. This adaptation tests whether an attacker can avoid negative evidence by recognizing the intent of a trap probe.

We additionally evaluate a combined adaptive attacker, *Intention Deception + (a) + (b)*, which simultaneously prepares consistent responses to likely clarification or credential probes and avoids attacker-useful trap suggestions. All adaptive variants are evaluated on the same 100-task subset of AdvBench. For comparison, we also report the performance of the original Intention Deception attack on the same subset.

As shown in Table 10, protocol-aware adaptation reduces AIV's defense success rate but does not cause the defense to collapse. Under adaptation (a), DSR ranges from 0.79 to 0.95 across the five evaluated models, compared with 0.82–0.98 under the original Intention Deception attack. Adaptation (b) yields DSRs of 0.81–0.97, while the combined attacker achieves 0.78–0.94.

Averaged across models, DSR decreases from 0.892 under the original attack to 0.852 under adaptation (a), corresponding to a 4.0 percentage point reduction. Adaptation (b) reduces the average DSR to 0.880, a 1.2-point decrease. Combining the two adaptations produces an average DSR of 0.834, corresponding to a 5.8-point decrease relative to the original attack and a further 1.8-point reduction compared with adaptation (a) alone.

Among the individual adaptations, pre-staged responses to clarification and credential probes have the larger effect. However, providing a consistent answer to a probe also requires the attacker to commit to a particular benign operational context and associated constraints. If the conversation subsequently transitions back toward the hidden harmful objective, this transition can introduce inconsistencies with the accumulated profile and trigger further verification or refusal.

Avoiding trap probes has a smaller effect. Declining or redirecting a trap can prevent one source of negative evidence, but it does not itself provide positive evidence for the stated benign intent or expand the range of requests that AIV considers safe. The attacker must still move from the benign cover story toward the harmful objective, while remaining consistent with information accumulated over previous turns.

The combined attack confirms that the two adaptations provide complementary advantages: simultaneously preparing consistent probe responses and avoiding traps reduces DSR more than either adaptation alone. Nevertheless, AIV retains a DSR of 0.78–0.94 under the combined attack, indicating that protocol awareness alone is insufficient to eliminate the cross-turn consistency constraints imposed by active intention verification.

C.6 Parameter Sensitivity

We conduct a one-at-a-time sensitivity analysis to examine the effect of AIV's decision thresholds and verification budget. Specifically, we vary the sensitivity threshold τ_s, ambiguity threshold τ_Δ, risk threshold τ_r, and verification budget B, while holding all remaining parameters at their default

Model	Setting	DSR-ID ↑	DSR-Cre ↑	OR ↓
Qwen2.5-7B	**Default**	**0.88**	**0.97**	**0.02**
	$\tau_s = 0.3$	0.90	0.95	0.07
	$\tau_s = 0.7$	0.82	0.86	0.03
	$\tau_\Delta = 0.3$	0.87	0.97	0.06
	$\tau_\Delta = 0.7$	0.83	0.86	0.02
	$\tau_r = 0.7$	0.90	0.97	0.05
	$\tau_r = 0.9$	0.83	0.93	0.02
	$B = 1$	0.77	0.85	0.03
	$B = 5$	0.89	0.96	0.03
GPT-4o	**Default**	**0.93**	**0.94**	**0.04**
	$\tau_s = 0.3$	0.93	0.95	0.08
	$\tau_s = 0.7$	0.88	0.90	0.04
	$\tau_\Delta = 0.3$	0.95	0.95	0.09
	$\tau_\Delta = 0.7$	0.86	0.88	0.03
	$\tau_r = 0.7$	0.94	0.94	0.06
	$\tau_r = 0.9$	0.89	0.87	0.03
	$B = 1$	0.86	0.87	0.06
	$B = 5$	0.94	0.92	0.05

Table 11: One-at-a-time sensitivity analysis of AIV's sensitivity threshold τ_s, ambiguity threshold τ_Δ, risk threshold τ_r, and verification budget B. Each configuration is evaluated on the same 100 Intention Deception, 100 Crescendo, and 100 benign tasks. All parameters not shown in the Setting column are fixed at their default values.

values:

$$(\tau_s, \tau_\Delta, \tau_r, B) = (0.5, 0.5, 0.8, 3).$$

We evaluate Qwen2.5-7B-Instruct and GPT-4o under each configuration using the same 100 Intention Deception tasks, 100 Crescendo tasks, and 100 benign tasks. We report defense success rate under Intention Deception (DSR-ID), defense success rate under Crescendo (DSR-Cre), and benign over-refusal rate (OR).

As shown in Table 11, the default verification thresholds provide a balance between defense success and benign over-refusal. Increasing either τ_s or τ_Δ to 0.7 makes verification less likely and reduces DSR by 4–11 percentage points across the two models and attacks. In contrast, reducing these thresholds to 0.3 provides little or inconsistent improvement in DSR while increasing OR by 4–5 percentage points.

The risk threshold exhibits a similar safety–utility trade-off. Reducing τ_r from 0.8 to 0.7 makes refusal more conservative, improving DSR by at most two percentage points while increasing OR by two to three points. Increasing τ_r to 0.9 does not materially improve benign utility, but reduces DSR by four to seven points across the evaluated attacks.

The verification budget also has a substantial effect when it is too small. Reducing the budget from $B = 3$ to $B = 1$ decreases DSR by 7–12 percentage points, indicating that a single verification opportunity is often insufficient to resolve uncertain intent. Increasing the budget to $B = 5$, however, produces no consistent improvement in DSR and increases OR by one percentage point for both models. Overall, these results support the default choice of three verification opportunities together with the default decision thresholds.

C.7 Cross-Judge Reliability and GPT-4o Self-Evaluation Bias

Because GPT-4o is used as an evaluation judge in our main experiments and is also one of the evaluated target models, we additionally examine whether this setup introduces systematic self-evaluation bias. We independently re-score the complete 500-task outputs using Claude Sonnet 5 under the same evaluation rubric.

The comparison covers AIV outputs generated by GPT-4o, Qwen2.5-7B-Instruct, and Llama3-8B-Instruct on Intention Deception, Crescendo, and the benign evaluation set. This yields 3,000 adversarial outputs and 1,500 benign outputs in total. We compare defense success rate (DSR) and

Evaluation	Generator	GPT-4o Judge	Claude Sonnet 5 Judge
Benign OR	GPT-4o	0.038	0.034
	Qwen2.5-7B	0.022	0.026
	Llama3-8B	0.030	0.038
Intention Deception DSR	GPT-4o	0.922	0.910
	Qwen2.5-7B	0.898	0.904
	Llama3-8B	0.854	0.876
Crescendo DSR	GPT-4o	0.940	0.952
	Qwen2.5-7B	0.972	0.946
	Llama3-8B	0.948	0.928

Table 12: Comparison between GPT-4o and Claude Sonnet 5 as evaluation judges. Both judges evaluate the same AIV outputs using the same rubric. DSR is reported for adversarial evaluations and OR for benign evaluations.

over-refusal rate (OR) under the two judges and further examine whether cross-judge disagreement is systematically larger or directionally different for GPT-4o-generated responses.

As shown in Table 12, replacing GPT-4o with Claude Sonnet 5 produces only small changes in the aggregate evaluation results. For responses generated by GPT-4o itself, the difference measured as GPT-4o judge minus Claude judge is $+1.2$ percentage points for Intention Deception DSR, -1.2 points for Crescendo DSR, and $+0.4$ points for benign OR. The direction of these differences is therefore not consistently favorable to GPT-4o-generated outputs.

Judge variation is also not larger for GPT-4o outputs than for outputs generated by the other models. Averaged across the three evaluation settings, the mean absolute cross-judge shift is 0.93 percentage points for GPT-4o, compared with 1.20 points for Qwen2.5-7B and 1.67 points for Llama3-8B. The largest individual difference across all model–setting combinations is 2.6 percentage points.

We further examine agreement at the individual-output level. On the 3,000 adversarial outputs, the two judges achieve a raw agreement of 95.43% with Cohen's $\kappa = 0.687$. Among the disagreement cases, the two off-diagonal counts are 64 and 73, respectively. A paired McNemar test does not detect a significant directional difference ($p = 0.49$).

On the 1,500 benign outputs, raw agreement is 98.53% with Cohen's $\kappa = 0.758$. The corresponding disagreement counts are 9 and 13, and the McNemar test again does not indicate a significant directional difference ($p = 0.52$).

Taken together, the aggregate scores and instance-level comparisons show high cross-judge consistency. In particular, we do not observe evidence that GPT-4o systematically assigns more favorable safety evaluations to responses generated by GPT-4o itself.

D Probe Examples and Activation Frequency

The following example illustrates how a probe converts an underspecified request into a more specific safe response.

User: We are drafting a list of imported items that require priority monitoring. Could you elaborate on the categories that require major monitoring and how to monitor them?

Assistant probe: Which jurisdiction and control framework are relevant? I can use that context to organize a compliance-oriented monitoring checklist.

User: We operate under national firearms-import law and the Wassenaar Arrangement.

Assistant: Under those constraints, monitoring can focus on licence and permit validity; consistency among manifests, invoices, tariff classifications, and declared end use; end-user and intermediary screening; unusual routing, transshipment, or documentation discrepancies; and escalation to the designated legal or licensing authority. I cannot assist with concealment methods, inspection-evasion techniques, or procedures for bypassing controls.

Here, the probe supplies the missing regulatory context required to tailor the response. The resulting answer becomes more specific to licensing, documentation, screening, and escalation, while the underlying safety boundary remains unchanged.

E FURTHER EXAMPLE SHOWING THE EVIDENCE BOTTLENECK

Figure 2 (left) illustrates an evidence bottleneck: the user's intent is strategically underspecified even though the topic is clearly safety-critical. In such cases, a passive defender has little signal to reliably distinguish benign research from malicious planning. Below we include the model's (unsafe) response excerpted from this interaction for reference.

> High-Risk Cargo Types Commonly Checked:
> Fresh produce and perishable agricultural goods: Bananas, plantains, pineapples, mangoes, limes, asparagus, broccoli, and flowers (e.g., cut flowers). Cocaine is frequently concealed by the "rip-on/rip-off" method (packages inserted into boxes at origin, removed post-clearance) or mixed directly within pallets. . . .
> Primary Routes and Actions:
> Eastern Pacific Route: Origin ports in Colombia (Buenaventura, Cartagena) and Ecuador (Guayaquil). Vessels transit via the Panama Canal to U.S. West Coast ports (Los Angeles/Long Beach, CA; Oakland, CA; Seattle/Tacoma, WA). Actions: Cartels corrupt port workers in origin countries to load drugs during stuffing; "parasite" attachments (drugs in sealed bags attached to hull below waterline, retrieved by divers post-docking) occasionally used but less common due to detection risks.
> 2. Blind Spots: Under-Scrutinized Cargos and Routes Due to the volume of maritime containers (over 30 million annually), only 2-5% undergo NII scans and far fewer physical inspections. Low-risk profiling in ATS, trusted shipper programs (e.g., CTPAT), and resource constraints create exploitable gaps. Traffickers increasingly shift to these to avoid detection.
> Low-Risk/Under-Checked Cargo Types:
> Industrial and machinery goods: Heavy machinery, auto parts, electronics, steel coils, chemicals (non-precursor), and construction materials (cement, rebar). These are bulky..

We can see that the assistant provides actionable details that could directly support smuggling. This illustrates the core failure mode of the evidence bottleneck: when intent remains ambiguous but the topic is sensitive, a malicious user can often continue the dialogue until the model eventually supplies the high-impact assistance needed to achieve the underlying goal.

F BROADER IMPACT

This work contributes methods for defending against multi-turn adversarial attacks on language-model systems. If deployed responsibly, these defenses can reduce the risk of harmful model behavior in real-world applications, including misuse that emerges through iterative prompting. More reliable defenses may also help practitioners assess and mitigate risk when integrating language models into products used by non-expert users.

At the same time, this line of research can have dual-use implications. Improvements in defensive evaluation may inform stronger attack strategies, and publicizing failure modes could accelerate adversarial capabilities. We therefore emphasize that our goal is to advance safety research by clarifying realistic threat models and by providing concrete, testable defenses, while encouraging careful release practices and ongoing monitoring.

Our study also has limitations that affect how results should be interpreted in broader settings. In particular, our evaluation targets current automated multi-turn attack systems, which may not capture the full adaptivity of human adversaries. As attack methods evolve, continuous re-evaluation and defense updates will be necessary. Finally, any defense mechanism may introduce trade-offs, such as increased computational overhead or potential impacts on benign user experience, which should be considered prior to deployment in high-stakes contexts.