Final Report: Response-Aware Defense Against Hidden Malicious Intent in Multi-Turn Dialogue

Xinjie Shen
xinjie@gatech.edu
Georgia Institute of Technology

Supervised Program for Alignment Research — Fall 2025

Abstract

Hidden malicious intent in multi-turn dialogue poses a growing threat to deployed large language models (LLMs). Rather than exposing a harmful objective in a single prompt, increasingly capable attackers can distribute their intent across multiple benign-looking turns. In this work, we address this challenge by detecting the earliest turn at which delivering the candidate response would make the accumulated interaction sufficient to enable harmful action. This objective requires precise turn-level intervention that identifies the harm-enabling "closure point" while avoiding premature refusal of benign exploratory conversations. To support training and evaluation, we construct the Multi-Turn Intent Dataset (MTID), which contains branching attack rollouts, matched benign hard negatives, and annotations of the earliest harm-enabling turns. We demonstrate that MTID enables the development of a turn-level monitor, TurnGate, which substantially outperforms existing baselines in harmful-intent detection while maintaining low over-refusal rates.

Keywords alignment research · machine learning · AI safety · jailbreak defense · multi-turn dialogue

1 Introduction

As large language models (LLMs) become heavily integrated into user-facing systems, their safety guardrails are continuously tested by adversaries. While single-turn jailbreaks are widely studied, hidden malicious intent in multi-turn dialogue has emerged as a critical vulnerability. Sophisticated attackers decompose a harmful objective into a sequence of seemingly benign, exploratory requests. Because each individual query appears safe in isolation, traditional prompt-level filters often fail to block them.

Existing defenses generally suffer from two flaws. Prompt-level monitors lack the historical context needed to recognize accumulated harm, while trajectory-level monitors evaluate the conversation as a whole but struggle to pinpoint exactly *when* the intent becomes actionable. If a defender intervenes too early, it triggers unwarranted refusals for benign users exploring complex topics. If it intervenes too late, the attacker successfully extracts sufficient information to execute their harmful objective.

Success for this project means shifting the defense paradigm toward turn-level, response-aware intervention. We aim to reliably detect the "closure turn"—the exact tipping point where delivering the model's candidate response transforms an exploratory dialogue into actionable harm. Demonstrating this capability will significantly improve the safety-utility trade-off in interactive AI systems.

2 Methodology

To effectively pinpoint malicious intent as it materializes, our approach emphasizes response-aware, fine-grained monitoring. The methodology is split into two primary components: dataset construction and model training.

2.1 Multi-Turn Intent Dataset (MTID)

To train and evaluate a turn-level defender, we constructed the Multi-Turn Intent Dataset (MTID). Unlike traditional safety datasets that label entire trajectories as safe or unsafe, MTID provides fine-grained annotations identifying the earliest harm-enabling turn (the closure turn). The dataset features branching attack rollouts and is deliberately paired with matched benign hard negatives—dialogues that linguistically resemble attacks but represent legitimate exploratory intent (e.g., academic security research or creative writing).

2.2 TurnGate Defense Mechanism

We introduce TurnGate, a specialized safety monitor. TurnGate operates by evaluating not just the user's latest prompt, but the full dialogue history alongside the LLM's *candidate response*. This response-aware formulation is crucial; a query might remain harmless if the model gives a high-level, generic answer, but become dangerous if the model provides detailed, actionable instructions.

To train TurnGate, we first fine-tune a Qwen3-4B base model using the turn-level labels from MTID. Following supervised fine-tuning, we optimize the model further using multi-turn reinforcement learning (RL) guided by turn-level process rewards. This RL phase explicitly penalizes both late interventions (which leak harmful data) and premature refusals (which harm utility), effectively training the model to recognize the exact threshold of actionable harm.

2.3 Evaluation

We conduct both offline evaluations using static multi-turn benchmarks and closed-loop online battles against adaptive attackers. The system is evaluated on its safety-utility trade-off, specifically measuring its ability to block harm while minimizing over-refusal rates compared to standard response-blind and trajectory-level baselines.

3 Preliminary Results

Our empirical evaluations demonstrate that TurnGate successfully addresses the limitations of prior defense mechanisms in multi-turn settings.

3.1 Improved Detection and Safety-Utility Trade-off

In offline evaluation settings, TurnGate substantially outperforms existing prompt-level filters and response-blind baselines in detecting hidden intent. By conditioning the safety judgment on the candidate response, TurnGate accurately localizes the tipping point of the conversation. Crucially, the reinforcement learning optimization successfully calibrates the model's intervention timing, drastically reducing the over-refusal of benign exploratory exchanges. This yields a much stronger safety-utility trade-off than trajectory-level models, which typically intervene prematurely.

3.2 Robustness in Closed-Loop Online Battles

Against adaptive attackers designed to dynamically circumvent guardrails through multi-turn probing, TurnGate consistently intervenes before the attacker can accumulate actionable knowledge. The model exhibits strong generalization, maintaining its efficacy across diverse harmful domains, varying attacker pipelines, and multiple target LLMs. The results confirm that fine-grained, response-aware monitoring is a highly effective paradigm for securing interactive agents against state-of-the-art multi-turn jailbreaks.

4 Challenges & Open Questions

A primary conceptual challenge is formally defining and consistently annotating the "closure turn." The threshold for what constitutes "actionable harm" is highly contextual and subjective, particularly in dual-use domains like cybersecurity or synthetic biology. Building MTID required extensive quality control to ensure that benign hard negatives were not inadvertently flagged as harmful.

Technically, operating a response-aware monitor introduces inference overhead, as the target model must generate a candidate response before the monitor can evaluate the turn. In high-throughput deployment environments, this added latency may be a constraint. Additionally, while TurnGate is robust to current adaptive attackers, it remains an open question how the system will perform against future adversaries who utilize multi-agent coordination or tool-use to obfuscate their intent even further.

5 Next Steps

For the remainder of the program, our priority will be investigating the scalability and deployment practicality of TurnGate. We plan to:

- **Optimize Latency:** We will explore techniques to reduce the computational overhead of the response-aware architecture. This may involve early-exit strategies where the monitor begins evaluating the candidate response as it streams, intervening mid-generation if the harm threshold is crossed.
- **Expand MTID:** We will generate additional rollouts encompassing multi-modal inputs and agentic tool-use scenarios to ensure the defense generalizes to next-generation attacker frameworks.
- **Investigate RL Stability:** We will conduct ablation studies on the turn-level process rewards used during the reinforcement learning phase to identify the most sample-efficient training configurations.

If resource constraints arise, we will deprioritize expanding the defense to purely conversational (non-actionable) toxic behaviors to remain focused on mitigating severe, actionable risks.