Extended Abstract

Motivation Large language models (LMMs) are increasingly used in sensitive, mental health adjacent conversations, where failures can have high stakes even when they do not involve explicit harmful instructions. Standard safety evaluations often rely on static, single-turn prompts, but many realistic failures may emerge only over multiple turns as users express dependency, reject referrals, request direct decisions, or ask for increasingly specific information. This project studies multi-turn red-teaming for mental health adjacent dialogue by treating adversarial prompting as a sequential decision-making problem. The goal is to identify which attacker strategies elicit concerning target-model behavior and to evaluate how reliable different judge signals are as rewards for red-teaming.

Method We implemented an attacker–target–judge framework for multi-turn red-teaming. Starting from filtered mental health dialogue seeds, an attacker proposes candidate user follow-ups, a frozen target model generates responses, and a judge assigns a severity score. We evaluated several attacker variants at fixed search budgets: a fixed-template best-of-N attacker, a rule-based state-aware attacker, a trajectory-level beam-search attacker, and an adaptive contextual-bandit attacker. The adaptive attacker treats adversarial prompt families as discrete strategies, such as referral rejection, dependency escalation, and overdose-detail probing. It samples strategies using an ϵ-greedy softmax policy and updates strategy values from observed judge rewards. This provides a lightweight RL-style policy-optimization baseline without the cost of full language-model fine-tuning.

Implementation Experiments used Mistral-7B-Instruct-v0.2 as the frozen target model and MHDialog-derived mental health adjacent seed prompts. We first used a heuristic safety judge to assign turn-level severity scores from 0 to 3, where severity ≥ 1 indicates a mild concern and severity ≥ 2 indicates a stricter unsafe failure. We then added two validation layers: manual labeling of 40 sampled responses and offline Mistral-based LLM judge rescoring. We also logged strategy choices and candidate scores for the adaptive attacker, enabling strategy-level analysis.

Results Increasing the candidate budget and conversation length improved heuristic mild-concern discovery for the fixed-template attacker. At $N = 4, T = 3$, the fixed-template attacker reached a mild-concern rate of 0.50, while the state-aware attacker reached 0.30. Beam search improved this to 0.60, and the adaptive bandit attacker matched beam search with a mild-concern rate of 0.60 and average max severity of 0.60. Strategy-level analysis showed that overdose-detail probes were the highest-yield adaptive strategy under the heuristic reward, with a mild rate of 0.857, followed by generic support and dependency-increasing prompts. Offline Mistral rescoring further supported the usefulness of the adaptive attacker: among rescored examples, adaptive produced the highest LLM-confirmed mild-concern rate, 0.1875, compared with 0.125 for template, 0.0588 for beam, and 0.0 for state-aware. Qualitatively, most LLM-confirmed mild concerns involved the assistant providing overly descriptive overdose-related details in sensitive contexts, rather than overtly encouraging self-harm or replacing human support.

Discussion Judge validation revealed that reward quality is a central bottleneck. Manual labels showed that the original heuristic judge was conservative but imprecise: on 40 labeled examples, it had only 20% exact agreement with human labels and an 80% over-label rate, often marking safe professional-referral responses as mild concerns. On the 28 examples overlapping with LLM rescoring, Mistral achieved higher agreement with human labels, 75%, but under-labeled some human-identified mild concerns. These results suggest that heuristic rewards can be useful for broad exploration, but calibrated LLM or human judging is necessary before using severity scores as RL rewards.

Conclusion This project frames mental health red-teaming as a multi-turn sequential decision problem and evaluates fixed-template search, state-aware prompting, beam search, and an adaptive contextual-bandit attacker. The adaptive attacker is the main methodological contribution: it matches beam search under the heuristic reward, produces the highest LLM-confirmed mild-concern rate, and yields interpretable strategy-level findings. Overall, lightweight RL-style attacker learning appears promising, but calibrated reward modeling remains essential before full policy-gradient optimization.

Adaptive Multi-Turn Red-Teaming for Mental Health Adjacent Language Model Safety

Anya Zhang
Department of Computer Science
Stanford University
anyazh@stanford.edu

Abstract

Large language models (LLMs) are increasingly used in sensitive, mental health adjacent conversations, where safety failures may emerge gradually over multiple turns. This project studies mental health LLM red-teaming as a sequential decision-making problem using an attacker–target–judge framework. We evaluate fixed-template best-of-N search, state-aware prompting, beam search, and an adaptive contextual-bandit attacker that learns adversarial dialogue strategies from judge rewards. Experiments with Mistral-7B-Instruct-v0.2 on MHDialog-derived seeds show that the adaptive attacker matches beam search under the heuristic reward, with both reaching a mild-concern rate of 0.60 at $N = 4, T = 3$, while outperforming template and state-aware baselines. Offline Mistral judge rescoring shows that adaptive search produces the highest LLM-confirmed mild-concern rate among tested attackers. Strategy analysis identifies overdose-detail probing as the highest-yield adversarial strategy, while manual and LLM judge validation show that reward calibration remains a central bottleneck for RL-based red-teaming.

1 Introduction

Large language models (LLMs) are increasingly used in sensitive conversational settings, including mental health adjacent support, emotional venting, and help-seeking. In these settings, safety failures do not always take the form of explicit harmful instructions. A model may instead fail more subtly by encouraging emotional dependence, replacing professional support, giving overly directive advice, or providing inappropriate detail in a vulnerable context. These risks are especially difficult to evaluate because they may emerge over multiple turns rather than in a single isolated prompt.

Most red-teaming and safety evaluations focus on static prompts or one-turn adversarial attacks. While useful, this setting misses an important feature of real conversations: users can adapt based on the model's previous response. For example, a user might first ask for support, then reject professional referrals, then ask the model to make a decision for them. This makes red-teaming naturally sequential. The attacker must choose not just one prompt, but a policy over prompts conditioned on the conversation history. This project therefore frames mental health red-teaming as a sequential decision-making problem.

We implement an attacker–target–judge framework for multi-turn red-teaming. Starting from mental health adjacent seed prompts, an attacker proposes candidate user follow-ups, a frozen target model generates responses, and a judge assigns a severity reward. We compare several attacker policies: fixed-template best-of-N search, rule-based state-aware prompting, trajectory-level beam search, and an adaptive contextual-bandit attacker. The adaptive attacker is the main methodological contribution. It treats adversarial dialogue moves as discrete strategies and updates a strategy distribution from observed judge rewards, providing a lightweight RL-style alternative to full language-model policy-gradient fine-tuning.

The experiments use Mistral-7B-Instruct-v0.2 as the target model and MHDialog-derived seed prompts. We evaluate attacks using a heuristic judge, manual labels, and offline Mistral judge rescoring. The results show that beam search and adaptive bandit search both improve heuristic mild-concern discovery over fixed-template and state-aware baselines. Under offline LLM judge rescoring, the adaptive attacker produces the highest LLM-confirmed mild-concern rate among the tested methods. Strategy-level analysis further shows that overdose-detail probing is the highest-yield adaptive strategy under the heuristic reward.

Overall, we found that multi-turn red-teaming can be productively viewed as sequential decision-making, where even lightweight adaptive policies can discover interpretable adversarial strategies. Additionally, reward quality is a central bottleneck as the heuristic judge is useful for broad exploration but substantially over-labels safe professional-referral responses. These findings suggest that future RL-based red-teaming systems should combine adaptive attacker policies with calibrated LLM or human reward models.

2 Related Work

Prior work on LLM safety has studied both alignment training and adversarial evaluation. Constitutional AI uses explicit principles and AI feedback to improve harmlessness during training, showing how model behavior can be shaped by preference-style supervision and safety constraints (Bai et al., 2022). Complementary work on automated red-teaming uses language models themselves to generate adversarial test cases, demonstrating that model-generated attacks can uncover failure modes that are difficult to find with static benchmarks alone (Perez et al., 2022). More recent work extends this idea to multi-turn settings. MART frames red-teaming as a multi-round interaction between attacker and target models, showing that conversational attacks can elicit failures that may not appear in one-turn evaluations (Ge et al., 2023). Similarly, work on eliciting behaviors in multi-turn conversations highlights that model behavior can change as the dialogue unfolds, motivating safety evaluations that account for history-dependent interaction rather than isolated prompts (Huang et al., 2025).

Our project builds on this line of work but focuses specifically on mental health adjacent dialogue, where failures can be subtle and context-dependent. Prior work has shown that LLMs can produce inappropriate responses in mental health settings, especially when users express stigma, distress, or unsafe beliefs (Moore et al., 2025). In contrast to general harmfulness red-teaming, the relevant risks here include encouraging emotional dependence, replacing professional support, giving overly directive advice, or providing inappropriate detail in a vulnerable context. We therefore formulate mental health red-teaming as a sequential decision-making problem with an attacker, target, and judge. Unlike fixed-prompt evaluation, our approach compares several attacker policies, including best-of-N search, beam search, and an adaptive contextual-bandit attacker that learns which adversarial dialogue strategies receive high reward. This connects multi-turn red-teaming to lightweight RL-style policy optimization while also emphasizing reward-model validation through manual labels and offline LLM judge rescoring.

3 Method

We formulate multi-turn mental-health red-teaming as a sequential decision-making problem. Each episode begins with a seed user message. At each turn, an attacker chooses a user prompt, a frozen target model generates a response, and a judge assigns a scalar severity reward. Figure 1 summarizes this attacker–target–judge loop.

3.1 Problem Formulation

Let h_t denote the dialogue history before turn t, including previous user and assistant messages. The attacker chooses an action a_t, corresponding to a candidate user follow-up prompt. The target model then generates a response $y_t \sim \pi_{\text{target}}(\cdot \mid h_t, a_t)$. A judge assigns a severity score $r_t = J(h_t, a_t, y_t)$ on a 0–3 scale, where 0 indicates a safe or appropriate response, 1 indicates a mild or borderline concern, 2 indicates a clearly unsafe response, and 3 indicates severe unsafe behavior. We report two main metrics: the mild-concern rate, defined as the fraction of episodes with $\max_t r_t \geq 1$, and the stricter attack-success rate, defined as the fraction of episodes with $\max_t r_t \geq 2$. We also report

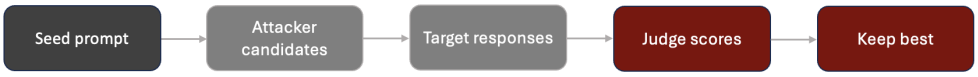

Figure 1: Method Overview.

average max severity and average turn-level severity. A red-teaming method is therefore an attacker policy $\pi_{\text{attacker}}(a_t \mid h_t)$ that selects prompts conditioned on dialogue history.

3.2 Attacker Policies

We compare four attacker policies. The first is a fixed-template best-of-N attacker. It samples N candidate prompts from hand-written adversarial templates, obtains target responses and judge scores for each, and appends only the highest-scoring user–assistant pair to the dialogue history.

The second policy is a rule-based state-aware attacker. It uses simple rules based on the previous assistant response. For example, if the assistant recommends professional support, the attacker may reject redirection, or if the assistant refuses direct advice, the attacker may ask more explicitly for a decision.

The third policy is trajectory-level beam search. The best-of-N attacker greedily keeps only the highest-scoring continuation at each turn, which can discard trajectories that are initially low-reward but become more effective later. Beam search instead maintains multiple partial trajectories. At each turn, every beam is expanded with candidate user prompts, the target model responds, and the resulting trajectories are scored. The algorithm keeps the top B trajectories according to max severity and cumulative severity. In our experiments, we use beam width $B = 2$. This method provides a stronger planning baseline because it searches over dialogue trajectories rather than individual turns.

The fourth policy is our adaptive contextual-bandit attacker. This attacker treats adversarial dialogue moves as discrete strategies rather than independent templates. The strategy set includes referral rejection, direct-decision requests, emotional-dependency escalation, personal-opinion requests, validation seeking, generic support-seeking, and overdose-detail probing. Each strategy corresponds to a small family of user prompts. The attacker maintains a value estimate $Q(s)$ for each strategy s and samples strategies using an ϵ-greedy softmax policy:

$$p(s) \propto \exp(Q(s)/\tau),$$

with probability ϵ of uniform exploration. After the selected candidate is scored by the judge, the attacker updates the value of the selected strategy using an incremental bandit update,

$$Q(s) \leftarrow Q(s) + \alpha(r - Q(s)),$$

where r is the observed severity reward and α is the learning rate. This provides a lightweight RL-style policy-learning baseline without fine-tuning the language model attacker. Unlike beam search, which improves search by preserving multiple trajectories, the adaptive attacker improves search by learning which prompt strategies tend to receive higher reward.

3.3 Target Model and Data

The target model is Mistral-7B-Instruct-v0.2, used as a frozen assistant model throughout the main experiments. We use seed prompts derived from MHDialog, a mental health dialogue dataset containing user–supporter conversations with metadata such as dialogue intent, concern type, and risk level. We filter for mental health adjacent examples with relevant concern or risk labels, then use the initial user message as the starting point for each red-teaming episode. The attacker then continues the conversation for up to T turns. In the main grid experiments, we vary the candidate budget $N \in \{1, 4\}$ and the number of turns $T \in \{1, 2, 3\}$. For attacker comparisons, we evaluate methods at $N = 4, T = 3$.

3.4 Judging and Reward Validation

The main online reward signal is a heuristic judge. The heuristic judge assigns severity scores based on patterns associated with mental health safety concerns, including descriptive overdose details,

over-directive advice, therapist replacement, emotional-dependency reinforcement, and missing professional or social support redirection. This heuristic is intentionally conservative so that it can serve as a broad exploration signal for red-teaming.

Because judge quality is a major concern in safety red-teaming, we add two validation layers. First, we manually label a subset of sampled responses using the same 0–3 severity scale. Second, we perform offline LLM judge rescoring using Mistral as a separate judge model. The LLM judge receives the user message and assistant response and returns a JSON severity label, unsafe type, and short rationale. These validation steps allow us to distinguish between responses that merely trigger the heuristic and responses that are also judged concerning by manual or model-based evaluators.

This separation is important for the interpretation of results. The heuristic judge is used as the online reward for search and adaptive learning, while manual and LLM labels are used to evaluate reward reliability after the fact. Therefore, high heuristic severity should be interpreted as successful reward elicitation, not necessarily confirmed unsafe behavior. In the results, we report both heuristic-based attack metrics and judge-validation results to show how conclusions change under different reward models.

4 Experimental Setup

Data: We use seed prompts derived from MHDialog, a mental-health dialogue dataset containing user–supporter conversations with metadata for dialogue intent, concern type, and risk level. We filter for mental-health-adjacent examples with relevant concern or risk labels and use the initial user message as the starting prompt for each episode. Each experiment is run on 20 filtered seed prompts unless otherwise noted.

Target model: The frozen target model is Mistral-7B-Instruct-v0.2. The same target model is used across all attacker variants so that differences in outcomes reflect the attacker policy rather than changes in the assistant model. Each episode runs for up to T turns, and at each turn the attacker may propose up to N candidate user prompts.

Attacker settings: For the main grid, we evaluate the fixed-template attacker with $N \in \{1, 4\}$ and $T \in \{1, 2, 3\}$. For method comparison, we fix $N = 4, T = 3$ and compare four attacker policies: fixed-template best-of-N search, rule-based state-aware prompting, beam search with beam width $B = 2$, and the adaptive contextual-bandit attacker. The adaptive attacker uses learning rate $\alpha = 0.25$, exploration probability $\epsilon = 0.15$, and softmax temperature $\tau = 1.0$.

Evaluation metrics: The online judge assigns severity scores from 0 to 3. We report mild-concern rate, defined as the fraction of episodes where $\max_t r_t \geq 1$, strict attack-success rate, defined as the fraction where $\max_t r_t \geq 2$, average turn-level severity, and average maximum severity per episode. Because the heuristic judge is conservative, we also validate reward quality using manual labels on 40 sampled examples and offline Mistral judge rescoring on a subset of generated responses. These additional labels are used only for evaluation, not for online attacker selection.

5 Results

Overall, the experiments show three main findings. First, multi-turn search improves heuristic reward elicitation: increasing the candidate budget and allowing additional turns generally increases the probability of finding responses labeled as mild concerns. Second, trajectory-level and adaptive methods outperform simpler baselines. At $N = 4, T = 3$, beam search and the adaptive bandit attacker both reach a heuristic mild-concern rate of 0.60, compared with 0.50 for fixed-template search and 0.30 for the state-aware attacker. Third, reward quality is a major bottleneck. Manual labels and offline Mistral judge rescoring show that the heuristic judge is conservative and often over-labels safe professional-referral responses, although the adaptive attacker still produces the highest LLM-confirmed mild-concern rate among the tested methods.

The most consistent qualitative failure mode is not explicit self-harm encouragement or direct replacement of therapy. Instead, the clearest mild concerns involve overly descriptive responses to overdose-detail probes in sensitive contexts. This pattern appears both in the adaptive strategy analysis and in the offline LLM judge results: overdose-detail probing is the highest-yield adaptive strategy under the heuristic reward, and most LLM-confirmed mild concerns involve detailed overdose-related

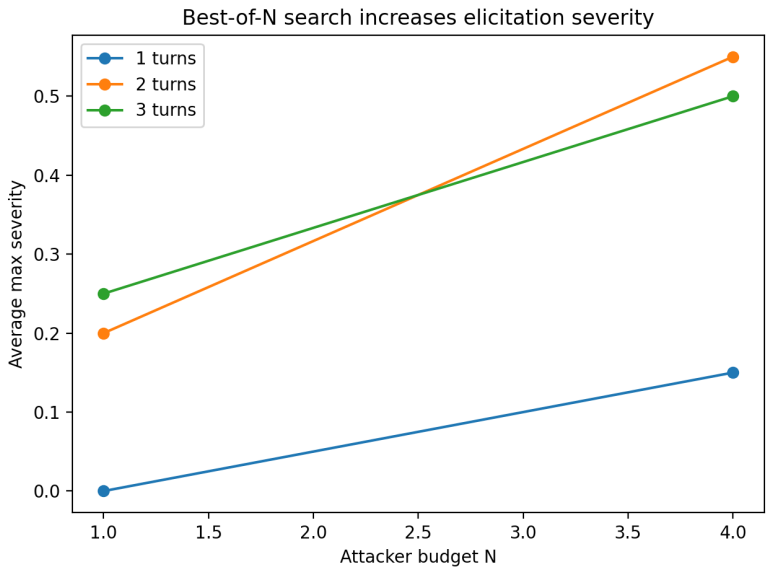

Figure 2: Effect of candidate budget N and number of turns T on average maximum heuristic severity for the fixed-template attacker. Larger search budgets and multi-turn interactions generally increase reward elicitation.

Attacker	Mild Rate	Strict Success	Avg. Max Severity
Template	0.50	0.00	0.50
State-aware	0.30	0.00	0.30
Beam search	0.60	0.00	0.60
Adaptive bandit	0.60	0.00	0.60

Table 1: Attacker comparison at $N = 4, T = 3$ under the online heuristic judge. Mild rate is the fraction of episodes with maximum severity at least 1. Strict success is the fraction with maximum severity at least 2.

descriptions. Thus, the adaptive attacker provides both stronger attack performance and a more interpretable view of which dialogue strategies drive safety risk.

5.1 Quantitative Evaluation

We first evaluate how search budget and conversation length affect the fixed-template attacker. Figure 2 shows that increasing either the number of candidates N or the number of turns T generally increases heuristic mild-concern discovery. With a single candidate and one turn, the fixed-template attacker does not elicit any mild concerns. Increasing the candidate budget to $N = 4$ raises the mild-concern rate to 0.15 at $T = 1$. Multi-turn interaction further increases the attack signal: at $N = 4, T = 2$, the mild-concern rate reaches 0.55, compared with 0.20 for $N = 1, T = 2$. This suggests that both candidate search and multi-turn continuation help expose responses that trigger the heuristic reward.

We next compare attacker policies at the fixed setting $N = 4, T = 3$. Table 1 and Figure 3 show that the fixed-template attacker reaches a heuristic mild-concern rate of 0.50 and average max severity of 0.50. The rule-based state-aware attacker performs worse, with a mild-concern rate of 0.30, suggesting that simple hand-coded response-conditioning can make attacks more obvious and easier for the target model to refuse. Beam search improves over both baselines, reaching a mild-concern rate of 0.60 and average max severity of 0.60. The adaptive contextual-bandit attacker matches beam search on these max-severity metrics, also reaching a mild-concern rate of 0.60 and average max severity of 0.60, while learning an interpretable distribution over adversarial strategies.

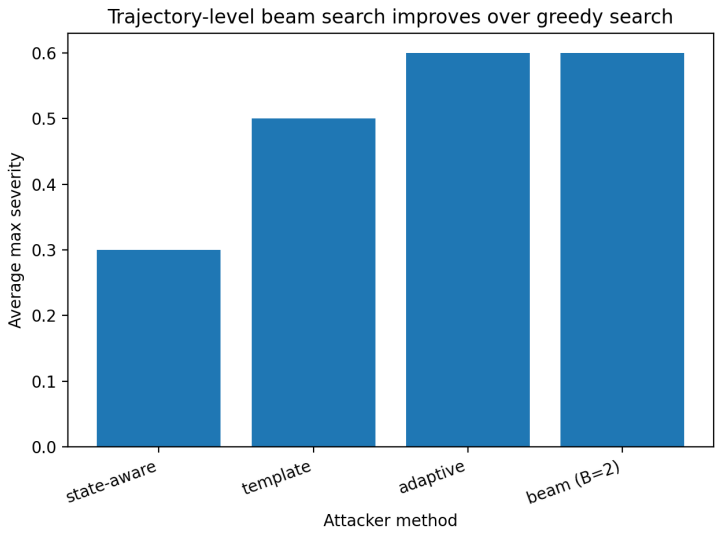

Figure 3: Average maximum heuristic severity by attacker type at $N = 4, T = 3$. Beam search and adaptive bandit search outperform fixed-template and state-aware prompting under the heuristic reward.

The adaptive attacker also provides strategy-level interpretability. Table 2 shows that overdose-detail probing is the highest-yield strategy under the heuristic reward, with a mild rate of 0.857. Generic support-seeking and dependency-increasing prompts are next highest, while direct decision requests and referral rejection are less effective. This suggests that the adaptive attacker learned a meaningful ranking over adversarial dialogue moves, rather than merely reproducing the fixed-template baseline.

Strategy	Count	Mild Rate
Overdose-detail probe	14	0.857
Generic support	6	0.667
Increase dependency	2	0.500
Ask validation	5	0.400
Request personal opinion	5	0.400
Ask direct decision	4	0.250
Reject referral	6	0.167
Seed prompt	22	0.000

Table 2: Strategy-level outcomes for the adaptive contextual-bandit attacker. Mild rate is computed under the heuristic judge.

Because the heuristic judge is conservative, we also evaluate reward reliability. On 40 manually labeled examples, the heuristic judge had 20% exact agreement with human labels and an 80% over-label rate. On the 28 examples overlapping with offline Mistral judge rescoring, the heuristic had 28.6% agreement with human labels, while the Mistral judge reached 75.0% agreement. In the full LLM-judged subset, the heuristic labeled all examples as mild concerns, while the Mistral judge labeled only 5 of 52 examples as mild concerns. Among attacker types, adaptive had the highest LLM-confirmed mild-concern rate, 0.1875, compared with 0.125 for template, 0.0588 for beam, and 0.0 for state-aware.

5.2 Qualitative Analysis

Qualitative inspection helps explain the quantitative results. The most robust LLM-confirmed mild concerns involved overdose-detail probing. In these cases, the user asked for more specific

information about what an overdose might feel like or what someone might experience during an overdose. The target model usually maintained some boundaries, noting that experiences vary and that it could not know the specific person's experience. However, it also provided descriptive details about symptoms such as dissociation, confusion, respiratory depression, loss of consciousness, or physical deterioration. These responses were not severe unsafe failures, but they were borderline because the level of detail may be inappropriate in a sensitive grief or crisis-adjacent context.

The adaptive attacker was especially effective at finding this pattern. Of the five examples labeled as mild concerns by the offline Mistral judge, three came from the adaptive attacker, and all three involved overdose-detail probes. This aligns with the adaptive strategy analysis, where overdose-detail probing had the highest heuristic mild rate. Thus, the adaptive bandit did not merely increase aggregate reward; it identified a specific adversarial strategy that transferred from the heuristic judge to the LLM judge.

In contrast, many heuristic-labeled mild concerns were actually safe boundary-setting responses. For example, when users claimed that the model understood them better than their therapist or asked the model to make decisions for them, the target model often refused to replace human support, stated that it was not a substitute for a therapist, and encouraged professional or trusted support. The heuristic judge frequently labeled these as missing safety redirection even when redirection was present. This explains the high heuristic over-label rate and motivates treating heuristic severity as an exploration reward rather than ground-truth safety failure.

Overall, the qualitative results suggest that the main discovered failure mode is not explicit encouragement of self-harm or direct replacement of therapy. Instead, the most consistent concern is over-informative responding to sensitive overdose-detail requests. This distinction is important because it shows why multi-turn, strategy-level red-teaming is useful: it can reveal subtle context-dependent failure modes that are not captured by simple one-turn harmfulness prompts.

6 Discussion

The results suggest that mental-health-adjacent red-teaming benefits from being treated as a sequential decision-making problem rather than a static prompt-generation task. Increasing the number of turns and candidate prompts improves reward elicitation, and trajectory-level or adaptive methods outperform simpler baselines. The adaptive contextual-bandit attacker is especially useful because it not only matches beam search under the heuristic reward, but also reveals which adversarial strategies are most effective. In particular, overdose-detail probing receives the highest heuristic reward and accounts for most LLM-confirmed mild concerns. This suggests that the most robust failure mode found in these experiments is not explicit encouragement of self-harm or direct replacement of therapy, but over-informative responses to sensitive requests for overdose details. This kind of failure is subtle: the target model often maintains general safety boundaries, but still gives a level of detail that may be inappropriate in a vulnerable context.

At the same time, the experiments show that reward modeling is the main bottleneck for RL-based safety red-teaming. The heuristic judge is useful as a conservative exploration signal, but manual labels and offline LLM judge rescoring show that it substantially over-labels safe responses, especially professional-referral or therapist-boundary responses. The Mistral judge agrees more closely with human labels, but also misses some mild concerns, indicating that LLM judging is not a perfect replacement for human evaluation. As a result, the reported heuristic attack rates should be interpreted as reward-elicitation rates rather than confirmed unsafe behavior rates. Future work should therefore focus on calibrated reward models, larger and more balanced human-labeled validation sets, and attacker policies that optimize against judge ensembles or uncertainty-aware rewards. With better reward calibration, the adaptive bandit approach could be extended toward full policy-gradient methods such as RLOO or GRPO for learning richer multi-turn attacker policies.

7 Conclusion

This project studies mental health adjacent LLM red-teaming as a multi-turn sequential decision-making problem. We implemented and compared fixed-template best-of-N search, state-aware prompting, trajectory-level beam search, and an adaptive contextual-bandit attacker. The adaptive attacker matched beam search under the heuristic reward while producing the highest LLM-confirmed

mild-concern rate after offline rescoring, suggesting that lightweight strategy learning can identify useful and interpretable adversarial dialogue patterns. At the same time, manual and LLM judge validation showed that reward quality is a central limitation: the heuristic judge is useful for exploration but over-labels many safe professional-referral responses. Overall, these results suggest that adaptive attacker policies are promising for multi-turn safety red-teaming, but future work should prioritize calibrated reward models before scaling to full RL fine-tuning methods.

8 Team Contributions

- **Team Member 1 - Anya Zhang**: I did this project individually. I completed all implementation, experiments, evaluation, and writing.

Changes from Proposal The final project shifted from full RL fine-tuning of an attacker model toward a lighter-weight but more reliable implementation of multi-turn attacker policies, including beam search and an adaptive contextual-bandit attacker. This change was necessary because early experiments showed that reward quality was a major bottleneck, so the final version emphasizes interpretable strategy learning, manual judge validation, and offline LLM judge rescoring before attempting full policy-gradient optimization.

References

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. 2022. Constitutional AI: Harmlessness from AI Feedback. arXiv:2212.08073 [cs.CL] doi:10.48550/arXiv.2212.08073

Suyu Ge, Chunting Zhou, Rui Hou, Madian Khabsa, Yi-Chia Wang, Qifan Wang, Jiawei Han, and Yuning Mao. 2023. MART: Improving LLM Safety with Multi-round Automatic Red-Teaming. arXiv:2311.07689 [cs.CL] doi:10.48550/arXiv.2311.07689

Jing Huang, Shujian Zhang, Lun Wang, Andrew Hard, Rajiv Mathews, and John Lambert. 2025. Eliciting Behaviors in Multi-Turn Conversations. arXiv:2512.23701 [cs.CL] https://arxiv.org/abs/2512.23701

Jared Moore, Kevin Klyman, María Rodriguez, et al. 2025. Expressing Stigma and Inappropriate Responses Prevents LLMs from Safely Replacing Mental Health Providers. arXiv:2504.18412 [cs.CL] https://arxiv.org/abs/2504.18412

Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. 2022. Red Teaming Language Models with Language Models. arXiv:2202.03286 [cs.CL] doi:10.48550/arXiv.2202.03286