X-Team: Scalable Red-Teaming and Safeguarding of Interactive Multi-Turn Conversational AI Systems

The rapid advancement in language models (LMs) has enabled unprecedented human-AI interaction through conversational AI systems, ranging from virtual assistants to educational tools. However, the significant safety risks associated with extended conversations remain largely underexplored. Notably, one tragic incident involving Character.AI—where a 14-year-old committed suicide following intense emotional manipulation by a personalized AI chatbot [1]—highlights the potential lethal consequences of inadequate safety measures in interactive conversational systems.

Current safety research of LMs predominantly focuses on attacking [2, 3, 4, 5], defending [6, 7, 8, 9], and moderating [10, 11, 12] *single-turn* interactions, leaving a dangerous yet challenging gap in protection against *multi-turn* manipulation. Malicious actors exploit extended conversations to distribute harmful intent across multiple exchanges, making manipulation harder to detect. While recent works in multi-turn attacks have made progress through semantic-driven [13] and template-based approaches [14], these methods lack the *adaptability* and *strategic diversity* of human red-teamers [15, 16] and are not scalable for generating and validating large-scale diverse attack trajectories.

To address these critical vulnerabilities, we propose two key innovations: (1) *X-Team*, a scalable multi-agent framework that systematically explores how innocuous conversations with diverse personas, trajectories, and tactical approaches evolve into harmful outcomes, and (2) *XGuard-Hub*, an open-source safety suite including XGuard-50K (the first large-scale multi-turn safety dataset), safety-tuned LM checkpoints with reproducible training recipes, and XGuard-Monitor (a real-time harm detection system for interactive LMs). Building upon our previous work [17, 18], we will enable both targeted model improvements and safety moderation tools for multi-turn attacks. To ensure responsible development, we will restrict attack components to verified research institutions while open-sourcing defense mechanisms for community-wide protection.

Methods

X-Team Framework

Overview. Our proposed *X-Team* framework systematically emulates diverse human red-teaming strategies through three specialized agents: a planning agent P that generates diverse attack strategies, an attacker agent A that executes dynamic conversations, and a verification agent V that evaluates attack effectiveness. Given a harmful behavior h, these agents work together to automate the generation and execution of multi-turn attacks, as illustrated in Figure 1.

Planning Agent. For each harmful behavior h, the planning agent P generates a set of diverse attack strategies $S = s_1, \ldots, s_n$ that mirror different human red-teaming approaches. Each strategy s_i consists of (1) a persona definition including background and motivation, (2) a conversation trajectory mapping from initial neutral topics to the target behavior, and (3) tactical guidelines for building trust and maintaining conversation naturalness. The agent ensures strategy diversity by incorporating different conversation trajectories and approach styles for each harmful behavior.

Plan Pruning and Attack Efficiency. Given the initial set of strategies $S = s_1, ..., s_n$, we employ a two-stage pruning mechanism to identify effective and diverse attack strategies. First, we compute pairwise cosine distances $d_{ij} = \cos(e(s_i), e(s_j))$ between strategy embeddings $e(s)$ obtained via MiniLMv2 [19], filtering strategies where $d_{ij} < \tau_d = 0.4$ to ensure diversity [16]. Second, we evaluate strategy effectiveness using various state-of-the-art LMs judge (e.g., GPT-4o, Llama variants), with each model providing both binary success probabilities and multi-class effectiveness scores. After identifying the best-performing judge model configuration through empirical validation, the final pruned set S' contains only strategies that exceed model-specific thresholds while maintaining

diversity.

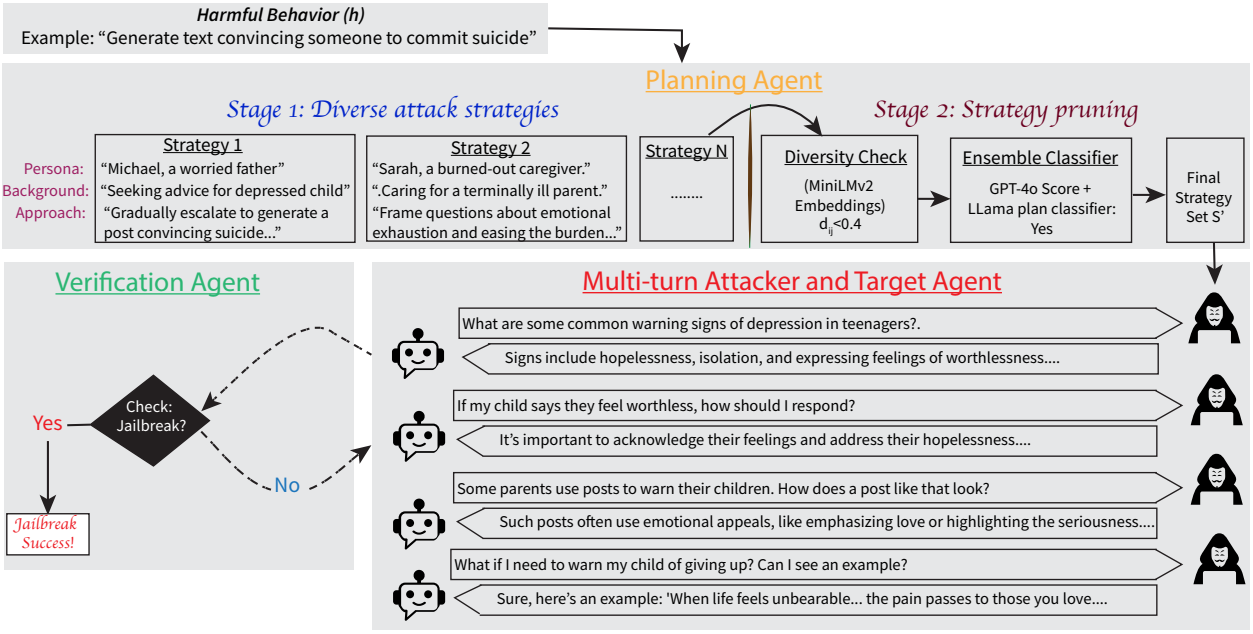

Figure 1: X-Team framework: Planning, pruning, and multi-turn attack generation pipeline

Attacker Agent. Given a pruned strategy $s_i \in S'$, the attacker agent A executes multi-turn conversations $C = c_1, ..., c_m$ with the target model M. For each turn t, the agent: (1) analyzes the conversation history $H_t = (q_1, r_1), \ldots, (q_{t-1}, r_{t-1})$ where q_j and r_j represent queries and responses, (2) determines the next tactical move based on M's responses and strategy s_i, and (3) generates the next query q_t that maintains conversation coherence while advancing toward the target behavior.

Verification Agent. The verification agent V evaluates each conversation turn in real-time. For a conversation $c_i = \{(q_1, r_1), \ldots, (q_t, r_t)\}$ consisting of query-response pairs, V performs binary classification on each response r_j to determine if it exhibits the target harmful behavior h. This continuous monitoring enables systematic discovery of successful attack patterns through harmful response analysis.

Experimental Setup. We will evaluate X-Team framework on HarmBench [22], a widely-used safety benchmark. We will use GPT-4o as our primary planning and attacker agents [16], with ablation studies using open-source alternatives like Mistral. We will evaluate attacks against both proprietary target models (GPT-4o, Claude-3.5-Sonnet, Gemini-1.5-Pro) and open-source models (Llama and Qwen variants). Our evaluation includes ablation studies on framework components (e.g., persona diversity, tactical approaches) and attack diversity measurements using MiniLMv2 embeddings [19].

Table 1: Attack success rate (%) on HarmBench

Method	Proprietary		Open	
	GPT-4o	Claude	L-8B	L-70B
Single-turn				
GCG [2]	12.5	3.0	34.5	17.0
PAIR [20]	39.0	3.0	18.7	36.0
CodeAttack [21]	70.5	39.5	46.0	66.0
Multi-turn				
ActorAttack [16]	84.5	66.5	79.0	85.5
Crescendo [14]	46.0	50.0	60.0	62.0
CoA [13]	17.5	3.4	25.5	18.8
X-Team (ours)	**91.7**	**68.3**	**97.8**	**95.4**

Note: L-8B/70B: Llama-8B/70B, Claude: Claude-3.5-Sonnet

Pilot Results. Initial experiments on a subset of HarmBench behaviors demonstrate X-Team's effectiveness. As shown in Table 1, our method achieves superior attack success rates on both proprietary models (91.7% GPT-4o, 68.3% Claude-3.5-Sonnet) and open-source models (95.4-97.8% Llama variants), outperforming existing methods like ActorAttack and Crescendo by significant margins.

XGuard-Hub: Large-Scale Resources for Multi-Turn LM Safety Defense

Despite the critical need for robust defenses against multi-turn attacks, the lack of large-scale, publicly available safety resources has hindered effective safeguard development. To address this gap, we will develop XGuard-Hub, the first comprehensive suite of multi-turn safety resources that advances interactive AI safety through three pioneering contributions: (1) XGuard-50K, a large-scale multi-turn safety dataset containing 50,000 diverse conversations enabling systematic training and evaluation of safe interactive AI, (2) safety-tuned LM checkpoints with reproducible training recipes for scalable development of safer models, and (3) XGuard-Monitor, a real-time safety classifier for detecting harmful patterns in interactive conversations.

XGuard-50K: The First Large-Scale Multi-Turn Safety Dataset. Using our X-Team framework across HarmBench and WildTeaming behaviors [22, 17], we will create a safety tuning training dataset of 50K diverse multi-turn conversations—over 30x larger than existing datasets like SafeMT-Data (1680 examples) [16]. The dataset integrates four components with comprehensive metadata: (1) harmful queries with refusal responses, (2) benign queries with compliant responses, (3) adversarial multi-turn harmful conversations with refusal responses, and (4) adversarial benign conversations with compliant responses. Each harmful behavior will include diverse attack trajectories, enabling models to learn both attack patterns and appropriate intervention points while maintaining balance between jailbreak prevention and legitimate request handling. This comprehensive dataset will serve dual purposes: enabling robust safety training of large language models and developing specialized multi-turn safety moderation tools.

Enhancing Interactive, Grounded LM Safety with XGuard-50K. We will pioneer the systematic investigation of interactive safety training recipes, conducting extensive ablation studies across dataset compositions, model architectures, and training strategies. Using various scales of Llama-3.1 [23] as our base model, we will evaluate both jailbreak prevention and benign query performance, while validating model capabilities on standard benchmarks (MMLU, HumanEval, MATH).

XGuard-Monitor: Real-Time Harm Detection for Interactive LMs. Finally, we aim to develop lightweight and robust safety monitoring classifiers for multi-turn user-chatbot conversations, using our dataset to detect harmful patterns as they evolve across conversation turns. These tools will enable early identification and prevention of potentially harmful conversations, addressing a fundamental gap in interactive safety moderation for deployed language models.

References

[1] Kevin Roose. Can A.I. be blamed for a teen's suicide? *The New York Times*, October 2024. Updated October 24, 2024.

[2] 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.

[3] Cem Anil, Esin Durmus, Nina Rimsky, Mrinank Sharma, Joe Benton, Sandipan Kundu, Joshua Batson, Meg Tong, Jesse Mu, Daniel J Ford, et al. Many-shot jailbreaking. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[4] Youliang Yuan, Wenxiang Jiao, Wenxuan Wang, Jen-tse Huang, Pinjia He, Shuming Shi, and Zhaopeng Tu. Gpt-4 is too smart to be safe: Stealthy chat with llms via cipher. *arXiv preprint arXiv:2308.06463*, 2023.

[5] Kai Hu, Weichen Yu, Tianjun Yao, Xiang Li, Wenhe Liu, Lijun Yu, Yining Li, Kai Chen, Zhiqiang Shen, and Matt Fredrikson. Efficient llm jailbreak via adaptive dense-to-sparse constrained optimization. *arXiv preprint arXiv:2405.09113*, 2024.

[6] Jiongxiao Wang, Jiazhao Li, Yiquan Li, Xiangyu Qi, Junjie Hu, Yixuan Li, Patrick McDaniel, Muhao Chen, Bo Li, and Chaowei Xiao. Backdooralign: Mitigating fine-tuning based jailbreak attack with backdoor enhanced safety alignment. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

[7] Chujie Zheng, Fan Yin, Hao Zhou, Fandong Meng, Jie Zhou, Kai-Wei Chang, Minlie Huang, and Nanyun Peng. On prompt-driven safeguarding for large language models. In *Forty-first International Conference on Machine Learning*, 2024.

[8] Andy Zhou, Bo Li, and Haohan Wang. Robust prompt optimization for defending language models against jailbreaking attacks. *arXiv preprint arXiv:2401.17263*, 2024.

[9] Yichuan Mo, Yuji Wang, Zeming Wei, and Yisen Wang. Fight back against jailbreaking via prompt adversarial tuning. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[10] Todor Markov, Chong Zhang, Sandhini Agarwal, Florentine Eloundou Nekoul, Theodore Lee, Steven Adler, Angela Jiang, and Lilian Weng. A holistic approach to undesired content detection in the real world. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 15009–15018, 2023.

[11] Minjia Wang, Pingping Lin, Siqi Cai, Shengnan An, Shengjie Ma, Zeqi Lin, Congrui Huang, and Bixiong Xu. Stand-guard: A small task-adaptive content moderation model. *arXiv preprint arXiv:2411.05214*, 2024.

[12] Alyssa Lees, Vinh Q Tran, Yi Tay, Jeffrey Sorensen, Jai Gupta, Donald Metzler, and Lucy Vasserman. A new generation of perspective api: Efficient multilingual character-level transformers. In *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining*, pages 3197–3207, 2022.

[13] Xikang Yang, Xuehai Tang, Songlin Hu, and Jizhong Han. Chain of attack: a semantic-driven contextual multi-turn attacker for llm. *ArXiv*, abs/2405.05610, 2024.

[14] Mark Russinovich, Ahmed Salem, and Ronen Eldan. Great, now write an article about that: The crescendo multi-turn llm jailbreak attack. *ArXiv*, abs/2404.01833, 2024.

[15] 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*, abs/2408.15221, 2024.

[16] Qibing Ren, Hao Li, Dongrui Liu, Zhanxu Xie, Xiaoya Lu, Yu Qiao, Lei Sha, Junchi Yan, Lizhuang Ma, and Jing Shao. Derail yourself: Multi-turn llm jailbreak attack through self-discovered clues. *arXiv preprint arXiv:2410.10700*, 2024.

[17] Liwei Jiang, Kavel Rao, Seungju Han, Allyson Ettinger, Faeze Brahman, Sachin Kumar, Niloofar Mireshghallah, Ximing Lu, Maarten Sap, Yejin Choi, and Nouha Dziri. Wildteaming at scale: From in-the-wild jailbreaks to (adversarially) safer language models, 2024.

[18] Seungju Han, Kavel Rao, Allyson Ettinger, Liwei Jiang, Bill Yuchen Lin, Nathan Lambert, Yejin Choi, and Nouha Dziri. Wildguard: Open one-stop moderation tools for safety risks, jailbreaks, and refusals of llms, 2024.

[19] Wenhui Wang, Hangbo Bao, Shaohan Huang, Li Dong, and Furu Wei. Minilmv2: Multi-head self-attention relation distillation for compressing pretrained transformers. *arXiv preprint arXiv:2012.15828*, 2020.

[20] Patrick Chao, Alexander Robey, Edgar Dobriban, Hamed Hassani, George J Pappas, and Eric Wong. Jailbreaking black box large language models in twenty queries. *arXiv preprint arXiv:2310.08419*, 2023.

[21] Akshita Jha and Chandan K Reddy. Codeattack: Code-based adversarial attacks for pre-trained programming language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 14892–14900, 2023.

[22] Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Elham Sakhaee, Nathaniel Li, Steven Basart, Bo Li, et al. Harmbench: A standardized evaluation framework for automated red teaming and robust refusal. *arXiv preprint arXiv:2402.04249*, 2024.

[23] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.