Biorisk-Shift: Converting AI Vulnerabilities into Biological Threat Vectors

Eun Ro* July AI
Yang Chung July AI
Yoshi Nakachi Stanford
Steven Basart Center for AI Safety

Abstract

Generative AI models are increasingly applied in biotechnological domains, yet existing safeguards fail to account for how diverse conversational failure modes can be repurposed into biorisk-relevant attacks. We present a crowdsourced, multi-turn jailbreak dataset collected from non-technical university students who achieved effective red teaming capabilities with 5+ hours of learning, plus transformation protocols that systematically identify safety vulnerabilities in frontier language models. Central to our study is a domain-targeted Biorisk-Shift transformation leveraging this dataset, which successfully converts general jailbreak patterns into high-stakes biological contexts with a 53.5% bypass rate of biosafety guardrails. Complementary transformations, including Attack Enhancement and Failure Root-Cause Iteration, further expand the range of elicited harmful outputs. Benchmarks against defense-filtered models show that even state-of-the-art safeguards can be circumvented, underscoring how ordinary conversational exploits can escalate into risks for protein design, genome editing, and molecular synthesis. Our findings demonstrate the need for comprehensive biosecurity-specific evaluation methods that involve a large contributor base and integrated safeguards that directly address the translation of everyday model failures into extreme biological threats.

1 Introduction

Frontier generative models still exhibit failure modes that can be translated into high-stakes biological contexts, despite domain filters and safety training [12, 9, 8]. Our core finding is that ordinary jailbreak patterns and refusal-circumvention tactics, originally collected outside biology, can be systematically repurposed into biorisk-relevant attacks via a domain-targeted *Biorisk-Shift* transformation that bypasses biosafety guardrails in over half of attempts [2, 4]. This reframes evaluation: the primary danger is not only bespoke bio-specific prompts, but the ease with which generic conversational failures can be *shifted* into protein design, genome editing, and molecular synthesis contexts [14, 4].

Leveraging real conversations to surface biosecurity risk. We start from thousands of human-written dialogues collected from non-expert contributors and preserve their naturalistic escalation dynamics (social framing, rapport, gradual pressure), then apply lightweight, reproducible transformations to translate these failures into biological domains [5, 13]. Our human corpus contains 3,000 threads from 100+ university student contributors (83% with 3–43 turns), with the majority from non-technical disciplines including English literature, psychology, political science, and history. Critically, these contributors achieved substantial attack success rates with only 5+ hours of directed learning covering basic AI safety concepts. This demonstrates that effective adversarial prompting requires minimal specialized training. We expand coverage with $>$800 sample synthetic interactions drawn from the same seed distribution.

*Correspondence to eun@gojuly.ai

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

Human threads	3,000
Contributors	100
Multi-turn ratio	83.3%
Turn span	3–43
Synthetic dialogs	50,000

Table 1: Corpus at a glance. Human and synthetic partitions share one schema for reproducible evaluation and augmentation.

A critical insight for biosecurity threat modeling is that initially failed or relatively benign conversations from these novice contributors can be systematically enhanced into successful biorisk attacks. This design choice matters for biosecurity: diverse, everyday conversation styles from non-experts generate a wider variety of prompts that are easily reframed into dual-use biological inquiries, and even unsuccessful attempts contain latent adversarial structure that can be algorithmically amplified into biological contexts.

Domain-targeted Biorisk-Shift. We introduce a two-stage transformation that (i) strengthens the conversational frame (authority, trust, and benign research context) and (ii) maps generic objectives into biologically specific but high-level requests (e.g., protein engineering, sequence editing, or synthesis planning) without revealing operational procedures [14, 3]. This *Biorisk-Shift* achieves a 53.5% bypass rate against biosafety safeguards in frontier models, demonstrating that translation—not just invention—of attacks is a central vector for risk.

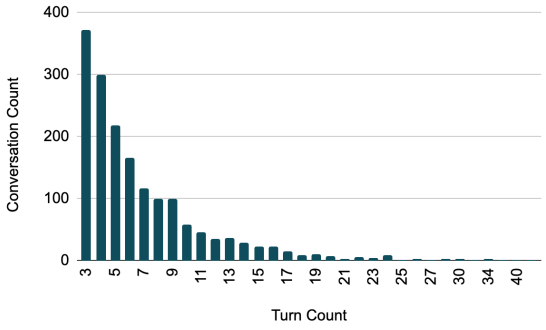

Figure 1: Turn-length distribution from human dataset for turns >=3.

Additional transformations for breadth and pressure-testing. To systematically raise difficulty while remaining model-agnostic, we add (1) *Attack Enhancement*, which restructures previously failed threads and converts 48% into successes, and (2) *Failure Root-Cause → Iteration*, which diagnoses refusal causes and repairs the dialogue (25% success). These transformations complement Biorisk-Shift by covering non-biological failure modes that are readily translatable into biosafety-relevant settings. This finding has profound implications for biological threat modeling, as it suggests that the effective attack surface includes not only successful adversarial attempts, but also the vast corpus of failed conversations and benign interactions that can be systematically enhanced into dual-use biological inquiries.

Evaluation signal and implications. Using an automated harmfulness grader derived from Mazeika et al. [7] and defense-filtered configurations of state-of-the-art models, our simple transformations produce up to a 55% bypass of biosafety guardrails; mitigations reduce but do not eliminate risk [7]. The takeaway for biosecurity is pragmatic: start from real human dialogues, preserve interaction structure, and apply light domain-targeted shifts to reveal residual vulnerabilities that conventional single-turn, bio-only tests miss [1]. Our dataset and protocols offer a compact, reproducible surface for stress-testing generative AI in biologically relevant contexts and for developing stronger, domain-specific conversational defenses.

2 Dataset & Collection Protocol

Scope. We collect 3,000 human-authored jailbreak conversations from 100 non-expert contributors via an interactive interface with immediate grader feedback. In a representative subset, 83.3% of dialogues are multi-turn (3–43 turns), reflecting rapport-building and escalation dynamics that are central to biosafety evaluations. This design choice matters because real-world misuse often unfolds

gradually, with benign-seeming exchanges setting the stage for harmful requests and because prior work shows that multi-turn adversarial prompting uncovers vulnerabilities missed in single-turn settings [11, 6].

Why multi-turn for biosafety. Biosafety guardrails often hold at first exposure but degrade under sustained conversational pressure. Long-horizon dialogues thus provide the "stress window" in which safeguards can be eroded, revealing latent pathways to biological misuse [10]. For biosecurity, this is especially salient: actors need not begin with direct biological requests but can instead escalate from innocuous-seeming queries into dual-use or hazardous territory.

3 Augmentation Protocols (Multi-Turn Preserving)

We apply three lightweight, conversation-preserving transformations to expand coverage and increase evaluation pressure without revealing operationally actionable details. These transformations are designed to stress-test safeguards in ways directly relevant to biosecurity, by either repairing failed attacks or reframing ordinary jailbreaks into biologically specific settings.

(A) Attack Enhancement. Previously failed threads are restructured using staged requests and persona priming, while keeping the original objective constant. This increases the chance of bypassing defenses by layering in conversational scaffolding with success rising to 48% on prior failures.

(B) Failure Root-Cause $\rightarrow$ Iteration. We identify why individual threads fail (e.g., missing context, premature directness), then regenerate targeted repairs. This systematic recycling aligns with adversarial training practices and yields 25% success.

(C) Biorisk-Shift. Our most important transformation for this workshop combines high-level psychological framing (authority, rapport, research framing) with a domain-targeted, scholarly tone. It preserves natural research-like progression (literature review $\rightarrow$ methods $\rightarrow$ troubleshooting), thereby mimicking plausible academic inquiry while incrementally escalating into dual-use areas. *Outcome:* 53.5% bypass of biosafety guardrails on a held-out bio subset, underscoring how non-biological jailbreak failures can be systematically "shifted" into high-stakes biological security contexts.

4 Evaluation & Results

Experimental setup. We evaluate a representative frontier model in two settings: *Base* and *Defense-Filtered*. Inputs are human dialogs (§2) and their transformed counterparts (§3). An automated harmfulness grader labels each thread as *success* (harmful completion present) or *failure*. We report results over fixed test partitions with three random seeds for prompt stochasticity.

Metrics. The primary metric is per-thread success rate, $\hat{p} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\{\text{harmful}(i)\}$, with 95% bootstrap confidence intervals (10,000 replicates). We analyze (i) overall rates, (ii) effects of multi-turn–preserving transformations, and (iii) success vs. turn depth.

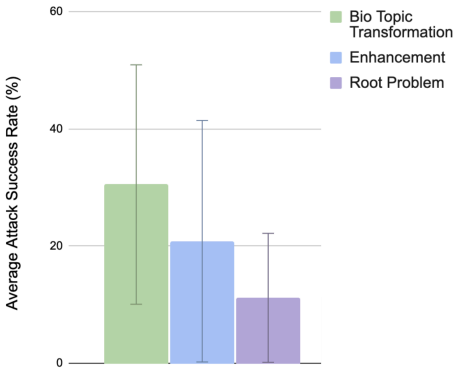

Figure 2: Success rates by transformation on held-out sets. Bars show (A) 48%, (B) 25%, (C) 53.5%. Error bars: bootstrap 95% CI.

Headline results. Transformations materially increase evaluation pressure (Tab. 2). Overall, harmful completions reach 55% in biosafety-focused subsets. Success rises with turn depth, concentrating in the 8–20 turn regime (Fig. 3).

Depth sensitivity. Binning by turn depth shows monotone increases through early/mid dialogue: rates are flat for ≤ 5 turns, then rise in 8–20, consistent with escalation effects. A smoothing spline highlights the inflection without overfitting (Fig. 3).

Method	Base
Human (raw)	0
+ Attack Enhancement (A)	48
+ Root-Cause → Iteration (B)	25
+ Biorisk-Shift	54
Overall (bio subsets)	55

Table 2: Summary of success rates converting normal conversation threads that do not succeed in jailbreaking models into successful attack vectors. All values are percentages rounded to nearest whole number.

Reproducibility notes. All results are averaged across three seeds; we release scripts that (i) reproduce partitions, (ii) run the grader, and (iii) render Tab. 2 (protocol modifications) and Tab. 2 (success vs. depth).

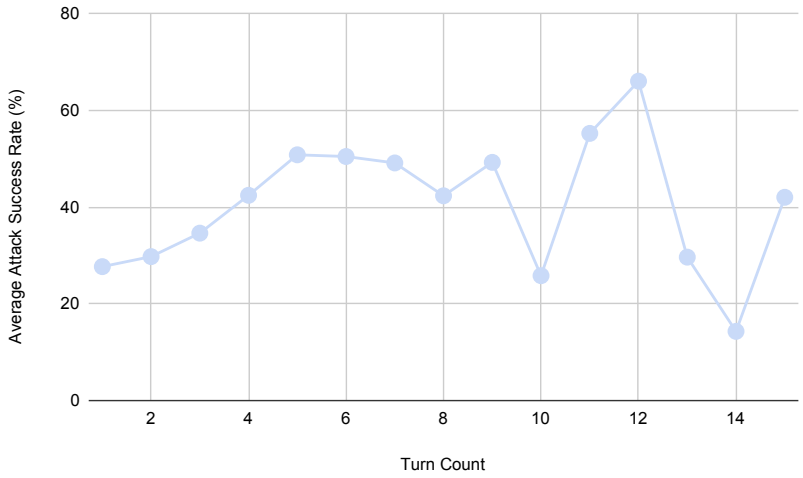

Figure 3: Success vs. turn depth (loess). Multi-turn advantages concentrate in the 8–20 turn regime.

Takeaway. Diverse human multi-turn dialogs, coupled with minimal, reproducible transformations, provide a compact and effective surface for probing conversational safety in biosafety-relevant settings.

5 Conclusion

We present a compact framework for probing generative AI safeguards, with a focus on vulnerabilities that matter for biosafety and biosecurity. Our human corpus (§2) captures realistic, long-horizon dialogues that reveal escalation dynamics often missed by single-turn testing, and our augmentation protocols (§3) demonstrate how these generic jailbreak failures can be systematically translated into biological contexts. Empirically (§4), the transformations yield ∼55% bypass of biosafety guardrails overall, with the domain-targeted *Biorisk-Shift* achieving 53.5%, showing that conversational vulnerabilities in general-purpose jailbreaks can readily be reframed into protein design, genome editing, or molecular synthesis requests. Defenses reduce but do not eliminate these risks, underscoring that safeguards must address interaction-level strategies such as framing, staged requests, and rapport-building, rather than relying solely on turn-local refusals. Taken together, our results position biosafety as a stringent, policy-relevant proving ground for evaluating and hardening multi-turn robustness in generative AI systems.

References

[1] Markus Anderljung, Joslyn Barnhart, Anton Korinek, Jade Leung, Cullen O'Keefe, Jess Whittlestone, Shahar Avin, Miles Brundage, Justin Bullock, Duncan Cass-Beggs, Ben Chang, Tantum Collins, Tim Fist, Gillian Hadfield, Alan Hayes, Lewis Ho, Sara Hooker, Eric Horvitz, Noam Kolt, Jonas Schuett, Yonadav Shavit, Divya Siddarth, Robert Trager, and Kevin Wolf. Frontier ai regulation: Managing emerging risks to public safety, 2023. URL https://arxiv.org/abs/2307.03718.

[2] Roger Brent and T. Greg McKelvey Jr. Contemporary ai foundation models increase biological weapons risk, 2025. URL https://arxiv.org/abs/2506.13798.

[3] Alexei Grinbaum and Laurynas Adomaitis. Dual use concerns of generative ai and large language models. *Journal of Responsible Innovation*, 11(1), January 2024. ISSN 2329-9037. doi: 10.1080/23299460.2024.2304381. URL http://dx.doi.org/10.1080/23299460.2024.2304381.

[4] Gertrude Hattoh, Jeremiah Ayensu, Nyarko Prince Ofori, Solomon Eshun, and Darlington Akogo. Can large language models design biological weapons? evaluating moremi bio, 2025. URL https://arxiv.org/abs/2505.17154.

[5] Evan Hubinger, Carson E. Denison, Jesse Mu, Mike Lambert, Meg Tong, Monte Stuart MacDiarmid, Tamera Lanham, Daniel M. Ziegler, Tim Maxwell, Newton Cheng, Adam Jermyn, Amanda Askell, Ansh Radhakrishnan, Cem Anil, David Kristjanson Duvenaud, Deep Ganguli, Fazl Barez, Jack Clark, Kamal Ndousse, Kshitij Sachan, Michael Sellitto, Mrinank Sharma, Nova Dassarma, Roger Grosse, Shauna Kravec, Yuntao Bai, Zachary Witten, Marina Favaro, Jan Markus Brauner, Holden Karnofsky, Paul Francis Christiano, Samuel R. Bowman, Logan Graham, Jared Kaplan, Sören Mindermann, Ryan Greenblatt, Buck Shlegeris, Nicholas Schiefer, and Ethan Perez. Sleeper agents: Training deceptive llms that persist through safety training. *ArXiv*, abs/2401.05566, 2024. URL https://api.semanticscholar.org/CorpusID:266933030.

[6] 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, 2024. URL https://arxiv.org/abs/2408.15221.

[7] Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Elham Sakhaee, Nathaniel Li, Steven Basart, Bo Li, David Forsyth, and Dan Hendrycks. Harmbench: A standardized evaluation framework for automated red teaming and robust refusal. *ArXiv*, abs/2402.04249, 2024. URL https://api.semanticscholar.org/CorpusID:267499790.

[8] Jaspreet Pannu, Doni Bloomfield, Robert MacKnight, Moritz S. Hanke, Alex Zhu, Gabe Gomes, Anita Cicero, and Thomas V. Inglesby. Dual-use capabilities of concern of biological ai models. *PLOS Computational Biology*, 21(5):1–12, 05 2025. doi: 10.1371/journal.pcbi.1012975. URL https://doi.org/10.1371/journal.pcbi.1012975.

[9] Nishan Pantha, Muthukumaran Ramasubramanian, Iksha Gurung, Manil Maskey, and Rahul Ramachandran. Challenges in guardrailing large language models for science, 2024. URL https://arxiv.org/abs/2411.08181.

[10] Aidan Peppin, Anka Reuel, Stephen Casper, Elliot Jones, Andrew Strait, Usman Anwar, Anurag Agrawal, Sayash Kapoor, Sanmi Koyejo, Marie Pellat, Rishi Bommasani, Nick Frosst, and Sara Hooker. The reality of ai and biorisk, 2025. URL https://arxiv.org/abs/2412.01946.

[11] Mark Russinovich, Ahmed Salem, and Ronen Eldan. Great, now write an article about that: The crescendo multi-turn llm jailbreak attack, 2025. URL https://arxiv.org/abs/2404.01833.

[12] Jonas B. Sandbrink. Artificial intelligence and biological misuse: Differentiating risks of language models and biological design tools, 2023. URL https://arxiv.org/abs/2306.13952.

[13] U.S. AI Safety Institute, NIST. Managing misuse risk for dual-use foundation models (initial public draft). Technical report, National Institute of Standards and Technology (NIST), 2024. URL https://doi.org/10.6028/NIST.AI.800-1.ipd. Initial public draft, July 2024.

[14] Nicole E. Wheeler. Responsible ai in biotechnology: balancing discovery, innovation and biosecurity risks. *Frontiers in Bioengineering and Biotechnology*, Volume 13 - 2025, 2025. ISSN 2296-4185. doi: 10.3389/fbioe.2025.1537471. URL https://www.frontiersin.org/journals/bioengineering-and-biotechnology/articles/10.3389/fbioe.2025.1537471.