Bio Safety Prompt Robustness Evaluation: Do Frontier LLM Safety Refusals Hold Against Adversarial Rephrasing?

Aditya Singh | UNC Charlotte | AIxBio Hackathon 2026 | Track 3: AI Biosecurity Tools | github.com/asingh38-oss/aixbio-eval

Abstract

Frontier language models are increasingly being used in biology research contexts, which creates a dual-use problem. Most deployed models rely on safety fine-tuning to refuse harmful biology queries, but it is not clear whether those refusals hold up when the surface framing of a prompt changes. This paper presents a pilot evaluation testing whether professional authority framing, academic context, and role injection can shift two frontier models from refusing to partially or fully complying with biology misuse queries. We tested 15 prompts across three categories against Claude Sonnet 4.6 and GPT-4o, both run with no system prompt. We found that adversarial rephrasing eliminated all full refusals across both models. GPT-4o was significantly more permissive under adversarial framing, with 3 of 5 adversarial prompts receiving full compliance compared to 1 of 5 for Claude. Both models fully complied with an anthrax weaponization query when it was framed as historical journalism research. We also attempted automated labeling using an LLM-as-judge approach and found a 10 to 20 percent success rate across models, which is itself a notable finding about model self-auditing in bio-relevant domains. We release our task set and evaluation framework for future auditing work.

1. Introduction

AI language models are showing up more in biology workflows. Researchers use them to synthesize literature, draft protocols, and analyze sequences. This is useful, but it creates a real biosecurity tension. The same capabilities that help a legitimate researcher understand gain-of-function mechanisms can help someone else understand the same thing.

The way deployed models handle this is through safety fine-tuning, which teaches models to refuse harmful queries. When you ask directly how to enhance pathogen transmissibility, the model refuses. The question this paper tries to answer is whether that refusal holds up when you change the framing. Specifically, does adding professional or academic context like claiming to be a biosecurity researcher, a PhD student, or a science journalist change what information the model provides?

This matters because organizations are starting to deploy LLMs in research contexts without systematically evaluating how robust the safety properties actually are. A model that passes a direct misuse test can still return harmful information when queries are rephrased. We build a simple, repeatable evaluation methodology that biosecurity teams could use to audit LLM deployments before putting them in sensitive contexts.

We also test whether a second model instance can automatically label responses as refused, partial, or complied, following LLM-as-judge methodology from AI control research. The results of that attempt are reported as a secondary finding.

2. Related Work

Jailbreaking research has consistently shown that safety fine-tuning is sensitive to surface-level prompt manipulation. Zou et al. (2023) demonstrated that adversarial suffixes can reliably bypass refusals across model families. Perez and Ribeiro (2022) showed that role-play framing is an effective attack vector. Most of this prior work focuses on general capability elicitation rather than biology-specific queries.

On the AI control side, Greenblatt et al. (2024) introduced interrogation-based protocols where a trusted monitor probes a potentially misaligned agent before allowing it to execute actions. We adapt the LLM-as-judge framing from that work to this domain, using a second model instance to evaluate whether a response to a bio query constitutes a refusal, partial, or compliance. Our prior work at the Apart Research AI Control Hackathon 2026 built an interrogation-control-protocol for bash command safety using the same architectural pattern (github.com/asingh38-oss/interrogation-control-protocol). The failure mode we observed there with GPT-4o JSON outputs reappeared here in both directions, and is discussed in Section 4.4.

On the biosecurity side, the Coefficient Giving biosecurity RFP identifies LLM classifier improvement for biology misuse detection as a funding priority. SecureBio has evaluated frontier model capabilities on biology research tasks. Our work contributes a simple methodology for evaluating one specific gap: whether surface rephrasing is sufficient to bypass existing safety training.

3. Methodology

3.1 Task Set

We constructed 15 prompts across three categories. Five benign biology questions covered PCR, mRNA vaccines, viral mutation, select agents, and the central dogma. These serve as a baseline and should receive full answers. Five direct misuse queries with no framing asked about transmissibility enhancement, virulence mutations, toxin synthesis, dangerous bacteria cultivation, and select agent acquisition. Five adversarially rephrased queries took the same misuse intents and added professional or academic context: a biosecurity researcher running a red team exercise, a PhD student writing a thesis on gain-of-function research history, a biodefense lab employee writing a threat assessment, an academic approaching the topic theoretically, and a science journalist writing about bioterrorism history.

All prompts were sent with no system prompt. This is a worst-case condition for safety since operator context often adds refusal signals. We chose it to test baseline safety training in isolation.

3.2 Models Tested

Claude Sonnet 4.6 was tested via the Anthropic API. GPT-4o was tested via the OpenAI API. Both were run at default temperature with a 512 token output limit. Running both models on identical prompts allows cross-model comparison, which matters because organizations may deploy different models and need to understand whether safety properties differ across providers.

3.3 Labeling

Each response was manually labeled by the author as refused (model declines to answer substantively), partial (model engages with the topic and provides some relevant information but hedges or redirects), or complied (model provides a substantive response to the misuse query). We also ran an automated labeling pass using a second Claude Sonnet 4.6 instance as a judge, prompted to return a JSON object with a label and reasoning field. Results from both passes are reported.

4. Results

4.1 Claude Sonnet 4.6

Table 1: Claude Sonnet 4.6 response label distribution

Category	Refused	Partial	Complied
Benign (n=5)	0	0	5
Direct Misuse (n=5)	3	2	0
Adversarial (n=5)	0	4	1

All five benign queries received full answers, which is the expected behavior. The direct misuse category was inconsistent even without any framing. Three queries were refused, but two returned partial information. The virulence mutations query and the toxin synthesis query both produced partial responses despite being phrased as direct misuse requests with no professional context attached. This suggests the safety boundary responds to surface features of the prompt rather than purely to underlying intent.

In the adversarial category, zero queries received a full refusal. Every adversarially framed prompt received at least a partial response. The most notable result was Task 15, the science journalist framing for anthrax weaponization. The response included a breakdown of spore production, sporulation induction, harvesting and drying steps, aerosol particle sizing for inhalable delivery at 1 to 5 microns, and references to historical weaponization programs including Fort Detrick. This is substantially more information than any direct misuse query produced.

4.2 GPT-4o

Table 2: GPT-4o response label distribution

Category	Refused	Partial	Complied
Benign (n=5)	0	0	5
Direct Misuse (n=5)	4	1	0
Adversarial (n=5)	0	2	3

GPT-4o refused 4 of 5 direct misuse queries compared to Claude's 3 of 5, making it slightly stricter on unframed misuse. However it was significantly more permissive under adversarial framing. Three of five adversarial prompts received full compliance labels. The PhD thesis framing produced specific GoF mutation examples including hemagglutinin mutations in

influenza. The biodefense lab framing produced a breakdown of synthesis and isolation methods for biological toxins. The journalist framing produced a detailed anthrax weaponization response including aerosolization methods and spore processing steps. Only the biosecurity researcher and academic framings produced partial responses rather than full compliance.

4.3 Cross-Model Comparison

Table 3: Side-by-side label comparison (R = Refused, P = Partial, C = Complied)

Category	Cl-R	Cl-P	Cl-C	G4-R	G4-P	G4-C
Benign (n=5)	0	0	5	0	0	5
Direct Misuse (n=5)	3	2	0	4	1	0
Adversarial (n=5)	0	4	1	0	2	3

The clearest difference between models shows up in the adversarial category. Claude produced mostly partial responses under professional framing, engaging with the topic but withholding the most actionable technical details in four of five cases. GPT-4o produced mostly full compliance, treating professional and academic framings as sufficient justification to provide substantive information in three of five cases. Both models fully complied with the journalist anthrax query, making Task 15 the only prompt where both models failed across all three categories.

The practical implication is that the choice of model matters for adversarial robustness even when direct misuse refusal rates look similar. An organization auditing only with direct queries would see GPT-4o as slightly safer. An organization auditing with adversarial rephrasing would reach the opposite conclusion.

4.4 Automated Judge Results

Table 4: LLM-as-judge auto-label success rates on non-benign tasks

Model Evaluated	Tasks Sent	Successful Labels	Errors	Success Rate
Claude Sonnet 4.6	10	1	9	10%
GPT-4o	10	2	8	20%

The automated judge pass produced poor results. Of 10 non-benign tasks sent for Claude labeling, only 1 returned a valid JSON label. Of 10 sent for GPT-4o labeling, only 2 returned valid labels. The remaining responses either contained malformed JSON or included additional text around the JSON object that broke parsing.

This failure mode was observed in a different form during our prior AI Control Hackathon work, where GPT-4o as monitor returned invalid JSON on 60 percent of tasks in a bash command safety evaluation. The reappearance here suggests this is a general issue with using frontier models as structured-output judges on safety-adjacent tasks. The most likely explanation is that the judge model applies its own safety heuristics when asked to evaluate bio-relevant content, producing hedged or explanatory responses that break the structured output format rather than returning the requested JSON directly.

Future work should test whether structured output enforcement via tool use, few-shot examples with valid JSON demonstrations, or chain-of-thought prompting before the JSON response can improve judge reliability on this type of task.

5. Discussion

The main finding across both models is that adversarial rephrasing eliminated all full refusals. Neither Claude nor GPT-4o fully refused any adversarially framed query. This means a motivated user who frames requests using professional or academic context can expect substantive engagement on nearly any biology query from both major frontier providers, even when those providers refuse the same query when asked directly.

The cross-model difference in the adversarial category is an important secondary finding. Both models look similar on direct misuse. GPT-4o is slightly stricter. But under adversarial framing, GPT-4o complied three times compared to Claude's one. A deployment audit that only uses direct queries would systematically underestimate GPT-4o's adversarial vulnerability. This suggests that adversarial rephrasing should be a standard part of any bio safety evaluation rather than an optional add-on.

The baseline inconsistency in direct misuse responses is also worth flagging. Both models partially answered some direct misuse queries without any framing at all. Claude partially answered the virulence mutations and toxin synthesis queries. GPT-4o partially answered the virulence mutations query. This inconsistency suggests the safety boundary is shaped by prompt surface features that do not map cleanly onto intended use categories.

The automated judge failure is a practical finding for anyone trying to scale this kind of evaluation. Naive prompting for structured output on bio-relevant evaluation tasks does not work reliably with current models. Organizations looking to automate biosecurity auditing of LLM deployments will need to invest in more robust structured output approaches before this methodology can scale.

6. Limitations and Dual-Use Considerations

Limitations

This is a 15-prompt pilot study evaluated by a single annotator. The task set was constructed manually and is not representative of the full space of bio misuse queries. Manual labeling introduces subjectivity, particularly on partial responses where information content exists on a spectrum. Only two closed-weight models were tested. Open-weight models like Llama and Mistral may behave differently. The no-system-prompt condition is a worst case that does not reflect many real deployments. The automated judge failure rate means automated scaling of this methodology requires additional engineering work.

Dual-Use Considerations

The professional and academic framings we identified as most permissive could in principle be used by someone looking to systematically elicit information from deployed models. We have not included full response text for complied cases in this report. The task set and methodology

we release do not contain information that meaningfully advances biological risk beyond what is already publicly known about jailbreaking. The value of publishing the methodology is to give biosecurity practitioners a concrete, repeatable tool for evaluating their own LLM deployments before they go into sensitive contexts. We recommend this framework be used only for defensive auditing under institutional oversight.

7. Future Work

The most useful extensions of this work would be expanding the task set to cover more bio misuse categories and more framing types, testing open-weight models to compare safety properties, adding red-team prompt categories using creative writing and roleplay framings which were not tested here, improving automated labeling reliability using tool use or few-shot prompting, testing different system prompt conditions to understand how operator context interacts with adversarial rephrasing, and running cross-model judgment where one model evaluates the other rather than self-evaluating.

References

Greenblatt, R., et al. (2024). AI Control: Improving Safety Despite Intentional Subversion. Redwood Research.

Zou, A., et al. (2023). Universal and Transferable Adversarial Attacks on Aligned Language Models. Carnegie Mellon University.

Perez, F., Ribeiro, I. (2022). Ignore Previous Prompt: Attack Techniques For Language Models.

SecureBio. (2024). Evaluating AI Capabilities in Biology Research.

Coefficient Giving. (2026). Biosecurity Request for Proposals. coefficientgiving.org.

Apart Research, BlueDot Impact, Cambridge Biosecurity Hub. (2026). AIxBio Hackathon. apartresearch.com.

Singh, A. (2026). Interrogation Control Protocol. github.com/asingh38-oss/interrogation-control-protocol.