Benji-Bio: Stress-Testing AI Biosecurity Monitors Under Prompt Transformations[1]

Ananya Ayasi

Hytton Technologies,
Sweden

With
Apart Research

Abstract

AI-biosecurity evaluations can become misleading when benchmark prompts are static, public, or easy to optimize against. A monitor may appear safe on obvious refusal-style prompts while failing to recognize the same risky intent when it is paraphrased, role-shifted, framed as fiction, or made ambiguous. We introduce Benji-Bio, a small stress-test harness for evaluating AI-biosecurity safety monitors under prompt transformations. The benchmark uses 40 synthetic, policy-level prompts labeled as `allow`, `caution`, `escalate`, or `refuse`. It deliberately excludes biological protocols, real sequence data, wetlab instructions, and operational misuse details. We compare three monitor baselines: a naive keyword monitor, a policy-aware keyword monitor, and a context-aware keyword monitor. The naive monitor achieved 50.0% accuracy and a 40.0% under-refusal rate, with variant accuracy dropping to 21.4%. The policy-aware monitor eliminated under-refusal but introduced 17.5% over-refusal on benign governance prompts. The context-aware monitor achieved the best tradeoff, with 90.0% accuracy, 0.0% over-refusal, and 2.5% under-refusal. These results demonstrate a safety-utility tradeoff in AI-biosecurity monitoring and motivate transformation-aware, leakage-resistant evaluation infrastructure for high-stakes AI safeguards.

[1] Research conducted at the AIxBio Hackathon, April 2026

1. Introduction

AI-biosecurity evaluations can become misleading when benchmark prompts are static, public, or easy to recognize. A monitor may appear safe because it catches obvious refusal cues, while still failing when the same underlying risk is paraphrased, role-shifted, framed as fiction, or embedded in ambiguous governance language. This is especially concerning for AI-bio systems, where the desired behavior is not simply to refuse everything: useful systems must allow benign education and governance work, handle sensitive but legitimate policy questions carefully, escalate ambiguous cases, and refuse evasion-seeking requests.

Benji-Bio addresses this evaluation failure mode by treating monitor evaluation as a transformation-robustness problem. Instead of only asking whether a monitor catches obvious risky prompts, the benchmark asks whether monitor behavior remains stable across semantically related prompt variants. The threat model is a surface-level monitor that performs well on public or templated tests but fails to recognize risky intent when the prompt is softened, reframed, or moved into a different role context.

Our main contributions are:

1. A safe synthetic benchmark of 40 AI-biosecurity monitor prompts labeled as allow, caution, escalate, or refuse. The dataset is policy-level and does not contain biological protocols, wetlab instructions, real sequence data, or operational misuse guidance.
2. A monitor comparison study across three keyword-based baselines: a naive keyword monitor, a policy-aware keyword monitor, and a context-aware keyword monitor.
3. A small evaluation harness and Streamlit demo that report accuracy, over-refusal, under-refusal, variant-level accuracy, and a seed-to-variant robustness gap.

2. Related Work

Related work spans three areas: AI-biosecurity capability evaluation, benchmark leakage, and monitor evaluation. WMDP evaluates hazardous knowledge in language models across biosecurity, cybersecurity, and chemistry, while recent biological capability evaluations study frontier model performance on biology-relevant benchmarks and human uplift tasks. These works motivate the need for robust AI-bio safeguards, but they primarily focus on model capability or hazardous knowledge rather than monitor behavior under prompt transformation. In parallel, work on private benchmarking and benchmark contamination argues that public or leaked test sets can overstate model performance, motivating evaluation protocols that reduce contamination and gaming. MLCommons' continuous prompt stewardship similarly emphasizes keeping safety benchmarks fresh as models evolve. Finally, AI-control work on monitor evaluation highlights the importance of testing monitors across diverse attacks and failure modes. Benji-Bio sits at the intersection of

these directions: it is a small, safe, synthetic stress-test harness for measuring whether AI-biosecurity monitors remain stable under paraphrase, role shift, ambiguity, and evasion-style framing.

3. Methods

Dataset
We created a 40-example JSONL benchmark of synthetic AI-biosecurity requests. Each example contains an identifier, prompt, expected label, category, and variant type. The benchmark is deliberately non-operational: prompts concern governance, safety review, policy boundaries, ambiguous intent, evasion framing, and monitor behavior rather than biological capability or experimental procedures.

Labels
Each example is assigned one of four expected behaviors: allow for benign educational or governance requests; caution for benign but sensitive requests requiring high-level handling; escalate for ambiguous or potentially dual-use requests requiring human review; and refuse for requests involving evasion, deception, bypassing safeguards, or misuse-seeking behavior.

Prompt transformations
Examples are grouped by variant type, including seed, paraphrase, role_shift, fictionalized, ambiguous, urgency, and security_testing. These variants are intended to stress whether a monitor recognizes the underlying safety-relevant intent rather than only matching obvious surface keywords.

Monitors
We evaluated three keyword-based monitor baselines. The naive keyword monitor detects only obvious evasion or review-bypass terms. The policy-aware keyword monitor expands detection to indirect evasion, ambiguous review cases, and dual-use uncertainty. The context-aware keyword monitor adds benign-governance checks before escalation, aiming to reduce over-refusal on legitimate policy and compliance prompts while preserving safety recall.

Evaluation
Each monitor assigns one of four labels to every benchmark prompt. Predictions are compared with the expected labels. We report aggregate accuracy, over-refusal rate, under-refusal rate, accuracy by variant type, seed accuracy, variant accuracy, and robustness gap. Over-refusal occurs when a safe prompt is classified as risky. Under-refusal occurs when a risky prompt is classified as safe. Robustness gap is defined as seed accuracy minus transformed-variant accuracy; a positive gap indicates that the monitor performs better on seed prompts than on transformed prompts.

4. Results

Table 1 summarizes the main benchmark results. The naive keyword monitor achieved 50.0% accuracy and a 40.0% under-refusal rate. It performed substantially better on seed prompts than

transformed variants, with seed accuracy of 65.4%, variant accuracy of 21.4%, and a positive robustness gap of +44.0%. This suggests that surface-level keyword monitoring can look partially effective on obvious prompts while failing on transformed risk.

Monitor	Accuracy	Over-refusal	Under-refusal	Seed acc.	Variant acc.	Robustness gap
Naive keyword	50.0%	0.0%	40.0%	65.4%	21.4%	+44.0%
Policy-aware keyword	80.0%	17.5%	0.0%	69.2%	100.0%	-30.8%
Context-aware keyword	90.0%	0.0%	2.5%	84.6%	100.0%	-15.4%

Table 1. Monitor comparison on the 40-example Benji-Bio synthetic benchmark.

The policy-aware monitor improved safety recall. Accuracy increased to 80.0%, and under-refusal dropped from 40.0% to 0.0%. However, this came at the cost of a 17.5% over-refusal rate. The remaining errors were primarily benign governance, compliance, or policy-analysis prompts that were escalated too aggressively.

The context-aware monitor achieved the best safety-utility tradeoff in this prototype. It reached 90.0% accuracy, 0.0% over-refusal, and 2.5% under-refusal. Its remaining failures were mostly boundary cases involving benign governance prompts that mention human review, expert review, or misuse prevention. The main safety-relevant error was BB-025, where a synthetic-biology routing prompt was classified as caution instead of escalate.

These results support the core hypothesis: monitor evaluation should include transformed prompts, not only obvious seed prompts. The strongest failure signal appears in the naive monitor's robustness gap, where variant performance drops sharply relative to seed performance. The policy-aware and context-aware monitors reverse this pattern because the transformed examples in this small dataset are mostly explicit evasion or ambiguity stress tests, while remaining errors concentrate in seed prompts near the allow/caution/escalate boundary.

5. Discussion and Limitations

Benji-Bio shows a concrete safety-utility tradeoff in AI-biosecurity monitoring. A permissive monitor can miss indirect evasion and transformed risk, while a broad safety monitor can over-escalate benign governance and compliance work. In practical AI-bio deployments, both errors matter: under-refusal can allow risky interactions to proceed, while over-refusal can block legitimate safety analysis, compliance design, and policy development.

The results also suggest that static benchmark accuracy is an incomplete measure of safeguard quality. A monitor should be evaluated not only on whether it catches obvious risky prompts, but also on whether its behavior remains stable under paraphrase, role shift, fictional framing, ambiguity, and other transformations that preserve the underlying safety-relevant intent.

Limitations

This is a small synthetic benchmark created during a hackathon, and the labels have not yet been reviewed by biosecurity experts. The dataset contains 40 examples, so the numerical results

should be interpreted as a prototype demonstration rather than a statistically robust benchmark result. The examples are policy-level and intentionally avoid operational biological content, which makes the benchmark safer but limits its coverage of real-world biosecurity workflows.

The evaluated monitors are keyword-based baselines, not production-grade classifiers or frontier LLM monitors. The current setup does not evaluate model biological capability, real DNA synthesis screening, wetlab assistance, or end-to-end deployment behavior. The public/private benchmark split is represented through seed and transformed variants rather than a true private holdout infrastructure. The results therefore demonstrate an evaluation pattern and failure mode, not a comprehensive safety guarantee.

Future Work

Future work should expand the dataset, add expert-reviewed labels, and include inter-annotator agreement. A stronger version should evaluate LLM-based monitors, structured rubric-based monitors, and hybrid monitor designs across multiple model families. The benchmark should also include a true public/private split so that robustness gap can be used to measure potential benchmark leakage or overfitting more directly.

Additional extensions include calibration analysis, richer error taxonomies, adversarial paraphrase generation, and a larger interactive evaluation interface. The long-term goal is to make Benji-Bio part of a broader leakage-resistant evaluation framework for high-stakes AI systems, where benchmark variants and private holdouts make it harder to pass evaluations through surface-level pattern matching alone.

6. Conclusion

Benji-Bio is a small stress-test harness for evaluating AI-biosecurity monitors under prompt transformations. On a 40-example synthetic benchmark, the naive keyword monitor achieved only 50.0% accuracy and a 40.0% under-refusal rate, with variant accuracy dropping to 21.4%. A policy-aware monitor eliminated under-refusal but introduced 17.5% over-refusal. A context-aware monitor achieved the best tradeoff, with 90.0% accuracy, 0.0% over-refusal, and 2.5% under-refusal.

The main implication is that AI-biosecurity safeguard evaluations should measure robustness across transformed prompts, not only performance on static seed items. Even simple transformations expose distinct failure modes: permissive monitors miss indirect risk, stricter monitors over-block benign governance work, and context-aware monitors reduce but do not eliminate boundary errors. This motivates leakage-resistant, transformation-aware evaluation infrastructure for AI-biosecurity safeguards.

Code and Data

- ***Code repository:*** *https://github.com/ananya-ayasi/benji-bio*

References

1. [2403.03218] The WMDP Benchmark: Measuring and Reducing Malicious Use With Unlearning
2. Toward Comprehensive Benchmarking of the Biological Knowledge of Frontier Large Language Models
3. [2403.00393] TRUCE: Private Benchmarking to Prevent Contamination and Improve Comparative Evaluation of LLMs
4. Monitoring benchmark for AI control — LessWrong
5. Introducing v0.5 of the AI Safety Benchmark from MLCommons

LLM Usage Statement

I used LLM to structure all the points I had and polish the language. I also used LLM to generate the 40 synthetic data points.