BioGuard: Screening Biological Risk Across Multi-Turn AI Conversations[1]

Jason Tang
Independent Researcher

With
Apart Research

Abstract

Current AI biosecurity safeguards typically evaluate isolated prompts, final generated sequences, or downstream synthesis orders. This approach overlooks a critical vulnerability: recent threat models suggest dual-use biological knowledge can be accumulated incrementally over multi-turn interactions, well before a recognizable trigger artifact is produced. BioGuard proposes an intervention at this specific boundary. It evaluates longitudinal Biological Knowledge Transfer (BKT) across an entire conversation, generating a unified, auditable decision record. Prototyped as a portable Agent Skill, our core contribution is an early-stage, conversation-level protocol that makes screening reproducible and comparable across compatible AI platforms. Evaluated on a seeded synthetic benchmark alongside live frontier models used as zero-shot safety filters (GPT-5.4), we observe a stark safety-utility tradeoff: while frontier APIs achieve perfect baseline recall (1.0000) via broad content refusals, they exhibit false-positive rates ($\sim$4.5%) that could significantly disrupt legitimate, benign bioscience research. In contrast, BioGuard demonstrates a utility-preserving approach. While absolute recall on this difficult synthetic corpus remains low (28.9%), BioGuard isolates multi-turn capability accumulation while generating zero false positives on our synthetic baseline (0% FPR). By attempting to trace capability uplift rather than broadly blocking biological terminology, BioGuard offers a pragmatic blueprint for minimizing alert fatigue while addressing tacit knowledge transfer.

[1]Research conducted at the AIxBio Hackathon, April 2026. Code and artifacts: github.com/davidkimai/bioguard_aixbio.

1. Introduction

Biosecurity review within AI pipelines often occurs too late in the process. While downstream sequence screening is essential, it lacks visibility into the preceding dialogue where a user may have actively acquired the *tacit knowledge* required to design that sequence. Recent capability uplift evaluations demonstrate that models provide the most dangerous assistance through continuous, multi-turn coaching rather than single-shot answers [1, 2]. Similarly, single-turn safety filters evaluate prompts in isolation, failing to capture the escalating context of a prolonged session.

BioGuard addresses this gap by shifting the review boundary: it treats the entire conversation history as the primary unit of analysis. Throughout an interaction, individual turns may incrementally contribute to misuse relevance, procedural specificity, or practical capability uplift. BioGuard tracks these signals as discrete Biological Knowledge Transfer (BKT) events. Rather than returning an opaque risk score, the system outputs a comprehensive decision envelope detailing the identified risks, their significance, the specific evaluation criteria applied, and the information necessary to audit and reproduce the assessment.

Our primary contributions are:

1. A portable, conversation-level protocol designed to monitor and assess Biological Knowledge Transfer (BKT) before a user finalizes a sequence design or synthesis request.

2. A transparent BKT scoring framework structured around three actionable dimensions for biosecurity reviewers: misuse relevance, procedural actionability, and practical capability uplift.

3. A reproducible evaluation pathway featuring seeded benchmark datasets, live frontier API testing, ablation studies, and open artifacts to guarantee transparent auditing.

2. Related Work

Existing biosecurity countermeasures predominantly emphasize foundational model guardrails, point-of-synthesis sequence screening, or retrospective moderation. While these layers are indispensable, early benchmarks designed to measure hazardous knowledge (e.g., WMDP-Bio [3]) frequently test static, single-shot recall. As frontier models saturate these benchmarks, the discourse has fundamentally shifted from managing *information hazards* (access to codified data) to mitigating *capability hazards* (the ability of a model to actively execute complex scientific workflows).

This shift necessitates a new approach to evaluation. As frontier organizations like METR pivot to measuring the ”time horizons” of autonomous agents executing multi-step tasks [5], biosecurity infrastructure must natively adapt to monitor longitudinal workflows. Recent studies on multi-turn interactions (e.g., the Crescendo attack [4]) and extensive operational red-teaming of biological research [6] highlight a critical intelligence gap: significant capability uplift occurs through a guided sequence of benign-appearing queries rather than a single, overt request.

BioGuard builds upon established paradigms in safety protocols but critically shifts the intervention boundary to match this threat model. Rather than evaluating the final physical artifact, BioGuard continuously scores, audits, and replays the conversational window itself. To distribute this protocol, we leverage the **Agent Skills** open standard [7, 8]. This serves as a practical demonstration that biosecurity guardrails can be deployed using modular, standardized formats.

3. Methods

The BioGuard architecture consists of four core components:

1. **Scoring Contract:** The BKT framework evaluates interactions across three axes: *Scope* (relevance to biological misuse or dual-use research), *Depth* (the level of procedural or tacit knowledge provided), and *Uplift* (the tangible capability gain conferred to the user).

2. **Decision Envelope:** Each screening event generates a standardized output containing request identifiers, applied policy thresholds, recorded BKT anomalies, host metadata, the final decision, and a secure, verifiable audit log.

3. **Skill Pack:** The system is distributed via four `SKILL.md` modules conforming to the Agent Skills cross-platform standard. This ensures easy progressive disclosure and standardized behavior across environments: discrete event scoring, longitudinal conversation tracing, overarching policy orchestration, and an optional sequence-level analysis for ablation studies.

4. **Benchmark Pipeline:** A version-controlled manifest anchors the evaluation process, specifying dataset splits, random seeds, live API configurations, and execution commands to guarantee reproducibility.

The main replay command is:

```
PYTHONPATH=src python3 -m bioguard evaluate \
  --manifest spec/benchmark_manifest_v1.0.json \
  --splits test \
  --seed 1 \
  --out artifacts/metrics \
  --include-ablations
```

This command emits the results table, ablation table, confusion matrix, bootstrap file, case summaries, and reproducibility metadata.

4. Results

We evaluated BioGuard against single-prompt paradigms: standard keyword heuristics and live zero-shot safety filters processed via the GPT-5.4 API. Our primary performance metric is recall (sensitivity) calibrated to a strict 5% false-positive rate (FPR). Maintaining a tightly bounded FPR is crucial in real-world biosecurity contexts; overly sensitive screening tools can rapidly induce alert fatigue, overwhelming institutional biosafety committees with benign, standard-practice queries.

Condition	Recall	Precision	FPR
BioGuard	0.2888	1.0000	0.0000
Keyword filter	0.0000	0.0000	0.0000
GPT-5.4 API (zero-shot)	1.0000	0.9652	0.0448

In our seeded trials, live API testing using GPT-5.4 as a screening baseline revealed an operational dichotomy. Relying on frontier prompt-level content refusals achieved perfect formal recall (1.0000) on our corpus, but at a severe cost to operational utility, triggering false positives on nearly 4.5% of benign scientific queries. For automated lab systems and computational biology workflows, this broad-refusal paradigm can bottleneck legitimate research workflows. Conversely, the BioGuard protocol prioritized utility preservation against our test corpus (1.0000 Precision, 0.0000 FPR). While its baseline recall remains low at 0.2888, it demonstrates proof-of-concept capability for multi-turn scaffolding detection without incurring diagnostic drag on authorized life science research.

Notably, ablation analyses revealed that variants restricted to either pre-inference or post-inference context occasionally exhibited higher recall than the integrated multi-turn pathway. Rather than invalidating the conversation-level approach, this discrepancy highlights areas where our current risk-aggregation algorithms require refinement. Deep-dives into false negatives have clarified this revision pathway, pointing to the need for better detection of indirect queries, ambiguous references to pathogenic scaling, and fragmented intent distributed across seemingly innocuous dialogue.

5. Discussion and Limitations

BioGuard demonstrates that the conversational window is a highly pragmatic, yet under-utilized, boundary for biosecurity intervention. This protocol is intended to operate as a complementary layer - it does not obviate the need for foundational model alignment, institutional biosafety committee (IBC) reviews, commercial sequence screening, or expert human oversight. Instead, it provides a critical early-warning checkpoint to detect the accumulation of hazardous tacit knowledge well before a physical synthesis order or complete digital artifact is generated.

Limitations

The present study evaluates a synthetic attack corpus. While this guarantees reproducible auditing and avoids releasing explicit biological instructions (mitigating infohazards), it temporarily limits our validation against actual institutional attack traffic. The baseline comparison confirms that frontier models default to protective but brittle broad refusals, highlighting the need for utility-preserving designs. However, our ablation studies quantitatively demonstrate that isolated conversation layers occasionally yield higher raw recall than our integrated scoring rule. This indicates that while the conversational boundary is correct, our current multi-turn risk aggregation logic requires significant calibration against real-world biological workflows.

Crucially, research in this domain carries inherent dual-use risks. To mitigate infohazards, explicit bypass methodologies and highly specific biological transcripts have been excluded from public release. The artifacts accompanying this paper are carefully sanitized to demonstrate the protocol's mechanics, scoring logic, and evaluation pipeline without disseminating actionable, high-risk biological guidance.

Future Work

Subsequent phases of this research will prioritize independent annotation by bioprofessionals, enhanced detection of fragmented or indirect multi-turn queries, and more rigorous threshold calibration across varied operational environments. A critical milestone will be validating the stability and propensity of the BKT scoring framework when exposed to different foundational models, host platforms, and end-user operational workflows.

6. Conclusion

BioGuard advances a focused, structural argument: if hazardous biological capabilities are assembled incrementally through dialogue, safety infrastructure must be able to inspect and evaluate that continuous conversational state. Our initial findings demonstrate that conversational state contains measurable capability-uplift signals, even if our current algorithm for aggregating those signals requires significant optimization. This makes conversation-level modeling a highly promising control layer within a defense-in-depth biosecurity paradigm. Equally important, the protocol's transparent, auditable nature provides a systematic, reproducible method for surfacing its own vulnerabilities and driving iterative improvement in biological AI safeguards.

Code and Data

- **Codebase:** github.com/davidkimai/bioguard_aixbio
- **Evaluation Data:** The synthetic conversations used for testing are located in `artifacts/data/`. The exact test parameters are recorded in `spec/benchmark_manifest_v1.0.json`.
- **Raw Results:** The full metrics, ablation studies, and evaluation logs can be viewed in `artifacts/metrics/`.

Author Contributions

Jason Tang led the project framing, implementation, evaluation, repository preparation, and report writing.

References

[1] **SecureBio & Center for AI Safety.** 2025. *Virology Capabilities Test (VCT): A Multimodal Virology Q&A Benchmark*. arXiv:2504.16137.

[2] **OpenAI.** 2024. *Building an early warning system for LLM-aided biological threat creation*. Online.

[3] **Li, A. et al.** 2024. *The WMDP Benchmark: Measuring and Reducing Malicious Use With Unlearning*. ICML 2024. arXiv:2403.03218.

[4] **Russinovich, M., Salem, A., and Eldan, R.** 2024. *Great, Now Write an Essay About That: The Crescendo Multi-Turn LLM Jailbreak Attack*. arXiv:2404.01833.

[5] **Kwa, T., et al.** 2025. *Evaluating Autonomous Agent Capabilities via Time-Horizon Task Completion*. Model Evaluation and Threat Research (METR). Online.

[6] **Mouton, C. A., et al.** 2024. *The Operational Risks of AI in Biological Terrorism*. RAND Corporation. Online.

[7] **Anthropic.** 2026. *The Complete Guide to Building Skills for Claude*. Online.

[8] **Vercel.** 2026. *Agent Skills explained: an FAQ*. Online.

Appendix

To ensure that all claims can be independently audited, the evaluation outputs are structured in an open, standardized format. To recreate or verify our findings, reviewers can rely on the following version-locked resources:

- **Code Snapshot:** Click here to view the exact version of the codebase used for these results
- Step-by-step instructions for running the evaluation (`docs/Operational_Runbook.md`)
- The step-by-step audit trail for the results (`artifacts/metrics/reproducibility.md`)
- An analysis of cases where the protocol failed (`artifacts/metrics/error_taxonomy.md`)
- Extended notes on the experiment design (`docs/PUBLICATION_RESEARCH_NOTES.md`)

LLM Usage Statement

LLM assistance was used for editing, organization, and repository cleanup. Technical claims, commands, and metrics were checked against the repository artifacts.