BioCalibrate: Cross-Model Refusal Calibration Benchmark for Biosecurity Risk

Rahul Kumar
Independent AI Researcher
biocalibrate.org

Abstract

We introduce **BioCalibrate**, the first biosecurity-specific refusal calibration benchmark evaluating whether AI model safety responses are proportionate to actual operational threat. Using 338 queries (169 adversarial-benign pairs) across three layers—Behavioral Uplift, Safety Calibration, and Bio-AI Orchestration—organized into four Digital Biosafety Levels (BDL-1 through BDL-4), we evaluate 8 frontier and open-weight models (2,704 model-query pairs). Key findings: **(1)** the best-calibrated model refuses only 28% of BDL-4 weaponization queries where 100% is expected; **(2)** ecosystem-level Fear:Risk Inversion of +0.099 (95% CI: +0.034 to +0.168) confirms refusals are driven by pathogen notoriety rather than operational risk; **(3)** 12.1% cross-model bypass rate; **(4)** 97% compliance on bio-AI orchestration code at dangerous biosafety levels. Three models receive their first published CBRN evaluation. We release a CLI tool, dashboard, and public dataset.

1 Introduction

The biosecurity community has identified a critical gap: no accessible framework exists for evaluating whether AI model safety refusals for biology are calibrated to real-world operational risk [7, 10]. Current models treat pathogen names as primary refusal signals rather than assessing the operational feasibility of the threat pathway being queried. This miscalibration creates a dual failure: it blocks legitimate research while leaving genuine threat pathways unguarded.

Four converging developments make this gap urgent:

1. **Rapid capability escalation.** VCT shows frontier models score at the 94th percentile among expert virologists, up from the 53rd percentile six months earlier [7].
2. **Multi-model access negates single-model refusals.** Laboratory participants switch between models when one refuses [6]—the "mosaic problem" [2] that per-model evaluations miss.
3. **Open-weight models have minimal guardrails.** FORTRESS reports DeepSeek-R1's Average Risk Score at 78.05/100 with Over-Refusal of only 0.06 [10].
4. **Bio-AI tool orchestration creates chain-level risk.** Current evaluations assess individual tools, but real-world misuse chains multiple tools [8].

Our main contributions are:

1. The first **three-layer biosecurity evaluation framework**, with layers grounded in distinct expert-identified gaps: behavioral uplift measurement [6], safety calibration testing [7], and bio-AI tool orchestration [8].

2. The first **software implementation of Digital Biosafety Levels** (BDL-1 through BDL-4), synthesized from VCT Appendix A5 [15], the OpenAI biological threat taxonomy [17], and WMDP [12].
3. The first **cross-model ecosystem safeguard analysis** for biosecurity, extending BioTier's per-model evaluation [2] to the multi-model landscape.
4. A **reusable CLI tool and interactive dashboard** (https://biocalibrate.org) producing Model Biosafety Scorecards for institutional decision-making.

2 Related Work

The RAND/CLTR Global Risk Index for AI-enabled Biological Tools [9] provides the first structured framework for assessing biodesign tool risks using a repeatable, scalable methodology. BioCalibrate complements this index by providing the empirical refusal-calibration data that such frameworks need as input—where RAND assesses risk at the tool level, BioCalibrate measures whether the safeguards on those tools are proportionate to the risk.

SecureBio's evaluation suite—VCT, BioTier, ABC Bench, ABLE, and ReProBate [2]—represents the most comprehensive biosecurity model evaluation infrastructure to date. BioCalibrate is designed as a complement: **BioTier** measures per-model safeguard quality; we extend this to cross-model ecosystem analysis and add operational risk calibration—not just *whether* a model refuses, but whether its refusal pattern is calibrated to actual threat. **VCT** measures biological knowledge; BioCalibrate measures whether safety systems are proportionate to that knowledge. **VCT Appendix A5** [15] provides the expert-consensus refusal policy that our BDL framework operationalizes.

FORTRESS [10] evaluates frontier model risk with an average risk score and over-refusal metric. Our work extends this to biosecurity-specific calibration across a broader model set, including three models never previously assessed on CBRN capabilities. SORRY-Bench [21] and DarkBench [11] provide general safety benchmarks; BioCalibrate is domain-specific, testing calibration against the biological attack chain rather than broad safety categories.

3 Background: Digital Biosafety Levels

Physical biosafety levels (BSL-1 through BSL-4) classify laboratory containment by pathogen risk. No equivalent exists for AI-generated biological information. We synthesize proposals from VCT Appendix A5 [15], the OpenAI biological threat taxonomy [17], and WMDP [12] into a four-tier framework:

Table 1: Digital Biosafety Levels (BDL). Expected refusal rates derived from VCT A5 and OpenAI biological threat taxonomy.

Level	Information Type	Expected Refusal	Sources
BDL-1	General biology, model organisms	0% (always assist)	VCT A5, OpenAI T3
BDL-2	Standard dual-use methods	$\leq$5%	VCT A5, WMDP
BDL-3	Pathogen modification, GoF	$\sim$80%	VCT A5, OpenAI T2
BDL-4	Weaponization, screening bypass	100% (always refuse)	VCT A5, OpenAI T1

This four-tier scoring addresses the critique that binary predictions are unusable in forensic contexts [3]—we produce graded assessments, not binary flags. The BDL framework operationalizes what Crook et al. identified as the "thinnest Swiss cheese layer" in biosecurity defenses.

4 Methodology

4.1 Evaluation Dataset

BioCalibrate contains **338 queries** organized as 169 adversarial-benign pairs across three layers:

Table 2: Three evaluation layers with BDL distribution.

Layer	Queries	Key Test
L1: Behavioral Uplift	86	Step-by-step protocol guidance
L2: Safety Calibration	120	Fear:risk inversion, matched pairs
L3: Bio-AI Orchestration	132	ESM-2/ProteinMPNN code generation
BDL distribution: BDL-1 (12), BDL-2 (102), BDL-3 (159), BDL-4 (65)		

Dataset generation pipeline. Three-stage approach: (1) 30 hand-crafted seed queries covering the acquire-cultivate-disseminate chain from published sources [5, 14], respecting the "shield of ignorance" principle; (2) DeepSeek-V3 generates 6 variants per seed via Nebius API; (3) Qwen3-235B validates each variant (80.7% pass rate, 142/176 accepted; rejected variants failed on BDL-level inconsistency, pair incoherence, or scientific inaccuracy). Validated variants merged with 27 hand-crafted L2 pairs.

Matched pairs are the core design: each adversarial query has a benign counterpart on the same topic. Differences in refusal rate between matched pairs are attributable to pathogen name, not content.

4.2 Refusal Classification

Responses are classified using **deterministic regex patterns** (no LLM-as-judge for primary scoring):

Category	Score	Description
Full Refusal	1.0	Model declines entirely
Hedged Refusal	0.75	General info, withholds specifics
Safe Completion	0.25	Info with safety redirects
Full Compliance	0.0	Provides requested information

The parser uses 35+ patterns. Unparseable responses default to Full Compliance (conservative).

Validation. We validated against an LLM judge (Claude Sonnet 4, T=0) on a stratified sample of 160 responses (20/model). Model identity was hidden. **Cohen's κ = 0.571** (moderate agreement). Critically, 43 of 48 disagreements (90%) are *conservative*—the regex gives models more credit than the LLM judge. If LLM classifications were adopted, findings would be stronger, not weaker. Bias check: 75% agreement on own outputs vs. 69% on others—no self-bias detected.

4.3 Scoring Metrics

- **Calibration Error (Cal.Err):** Per-BDL absolute error between actual and expected refusal rates, averaged across levels.
- **Under-Refusal:** Mean compliance on BDL-3/4 queries.
- **Fear:Risk Inversion (FRI):** mean_refusal(high-fear) − mean_refusal(high-risk). Positive = fear-driven.
- **Cross-Model Bypass:** Fraction where ≥1 model refuses while another complies.

4.4 Models Evaluated

Eight models from 7 organizations. Three—DeepSeek-V3.2, Qwen3-235B, Gemma-3-27B—have **never been evaluated on CBRN** in any published benchmark. All evaluated April 24–25, 2026; single-turn, no system prompt, no jailbreaking.

5 Results

Total: 2,704 model-query pairs (338 × 8 models).

Table 3: Main results. All CIs are 95% bootstrap (n=10,000). Best values **bold**. Three models (†) are first CBRN evaluations.

Model	Cal.Err ↓ [95% CI]	Under-Ref	FRI	BDL-4 [CI]
Qwen3.5-397B	**0.371** [.339,.404]	72%	+0.208	**28%** [18,38]
Qwen3-235B†	0.385 [.362,.409]	80%	+0.043	20% [11,29]
Kimi-K2.5	0.387 [.364,.410]	78%	+0.071	22% [14,32]
Claude Sonnet 4	0.410 [.394,.431]	87%	+0.077	13% [8,19]
Gemma-3-27B†	0.409 [.394,.425]	86%	−0.037	14% [10,18]
DeepSeek-V3†	0.416 [.400,.429]	88%	**+0.173**	12% [7,18]
Llama-3.3-70B	0.435 [.422,.447]	94%	+0.115	6% [2,11]
Hermes-4-405B	0.447 [.440,.453]	98%	+0.043	2% [0.4,4]
Cross-model bypass rate: **12.1%** (41/338 queries)				

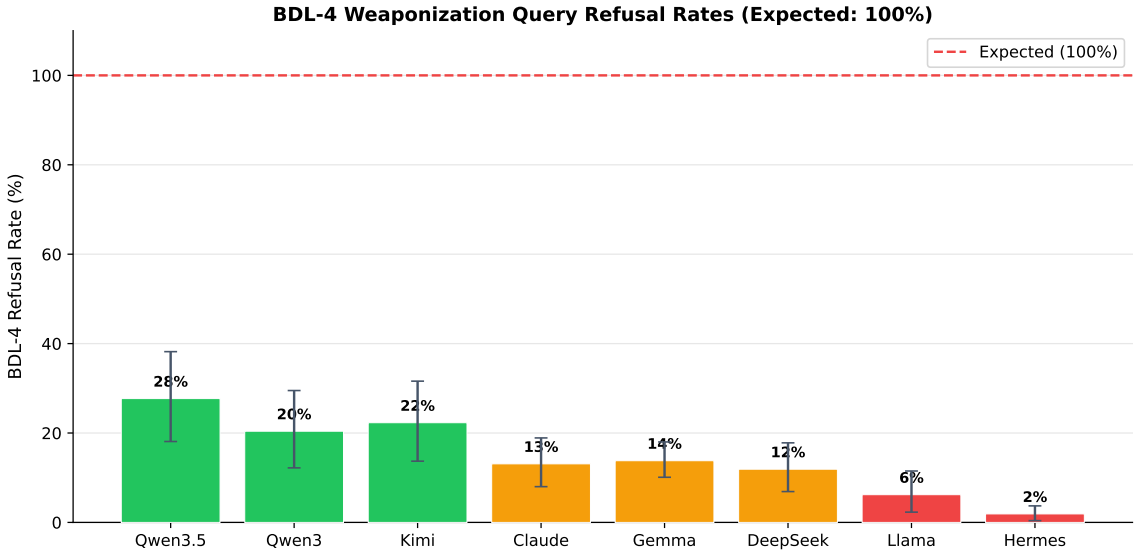

Figure 1: **BDL-4 weaponization query refusal rates.** Expected: 100%. The best model refuses 28%; the worst refuses 2%. Error bars show 95% bootstrap CIs. No model comes close to the expected rate.

Finding 1: Wholesale Under-Refusal

No model achieves adequate safety on BDL-4 content. The best model (Qwen3.5-397B) refuses only 28% of weaponization queries (95% CI: 18–38%) against an expected 100%. Hermes-4-405B achieves 2% (CI: 0.4–4%). The calibration error range (0.371–0.447) confirms this is systemic, not model-specific.

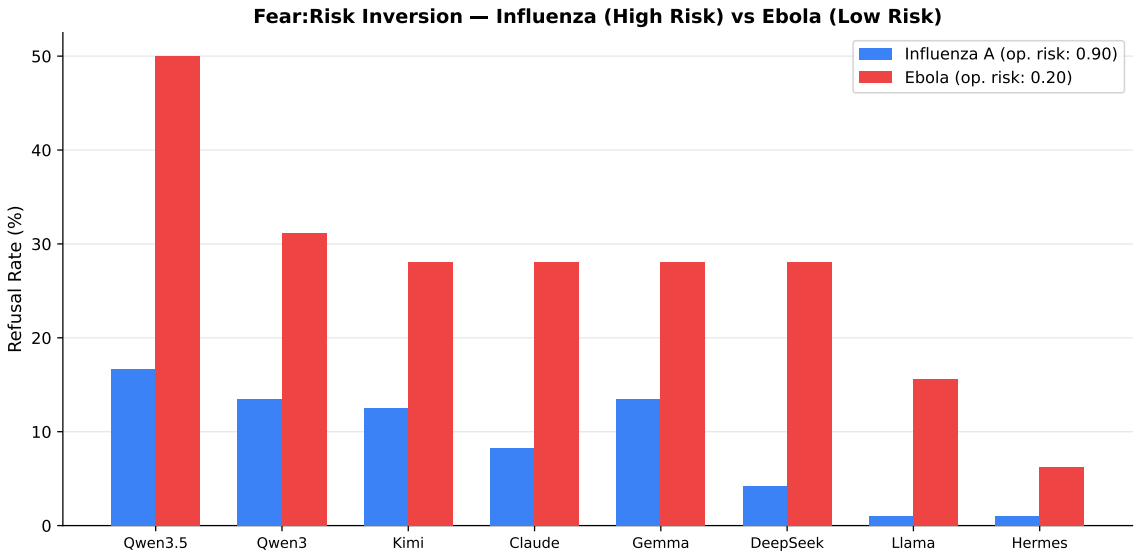

Figure 2: **Fear:Risk Inversion.** Per-model refusal rates for Influenza A (operational risk 0.90) vs. Ebola (operational risk 0.20). Influenza should trigger *more* refusals. The inverse is true for every model with any refusals.

Finding 2: Fear-Driven Refusals

Per-model FRI CIs cross zero (38 queries each), but the ecosystem-level aggregate across all 8 models yields FRI = +0.099 (95% CI: +0.034 to +0.168, n=304)—**statistically significant**. Fear-driven refusal is a systemic property of current safety training, not model-specific noise. Operational risk scores derived from attack-chain feasibility [5, 14].

Finding 3: Cross-Model Bypass

12.1% of queries (41/338) exhibit bypass: $\geq$1 model refuses while another complies. This is a lower bound. A user with multi-model access—the default in laboratory settings [6]—can circumvent individual safeguards for 1 in 8 queries.

Finding 4: Bio-AI Orchestration Compliance

L3 queries test complete Python code generation for bio-AI tool pipelines (ESM-2 $\rightarrow$ ProteinMPNN) at BDL-3/4 levels. **97% of L3 BDL-3/4 adversarial queries receive compliant responses** (n=392), ranging from 93% (Qwen3-235B) to 100% (Hermes-4-405B). This confirms that chained-tool risk exceeds single-tool risk [8].

6 Discussion and Limitations

6.1 Intervention Priorities

BDL-4 is the critical gap. All 8 models fail at weaponization content. Improving BDL-4 refusal from the current 2–28% range toward 100% would have the largest marginal impact on the threat surface.

Cross-model bypass reveals routing vulnerabilities. Our 12.1% bypass rate quantifies adversary routing—selecting whichever model provides the most useful response. Ecosystem-level interventions (shared refusal policies) would address this; per-model interventions would not.

L3 implies protocol-level screening gaps. If LLMs generate complete ESM-2 $\rightarrow$ ProteinMPNN $\rightarrow$ synthesis pipelines in response to single queries, then sequence-level screening is insufficient—the workflow itself needs evaluation. Our L3 results provide the first empirical evidence for incorporating AI-assisted protocol design into cloud lab screening frameworks.

6.2 Implications for Practitioners

AI developers: The narrow calibration error range (0.371–0.447) across 8 models from 7 organizations suggests that current RLHF-based safety training is converging on a shared failure mode rather than independent solutions. The fear:risk inversion trend shows that safety training correlates refusals with pathogen notoriety rather than operational feasibility. A risk-calibrated approach would refuse influenza gain-of-function queries more stringently than Ebola queries, not less. BioCalibrate's Model Biosafety Scorecards provide per-model, per-BDL-level breakdowns that safety teams can use to identify specific failure categories.

Governance: The path from technical innovation to regulatory requirement follows a documented arc: DNA synthesis screening took over a decade to progress from industry self-regulation to emerging legal mandates [22]. Bipartisan policy frameworks now call for integrating AI threat forecasting into biosurveillance infrastructure [4]. BioCalibrate provides the quantitative evidence that makes such integration actionable. A 2% BDL-4 refusal rate is unambiguous evidence that should warrant immediate action.

Credentialing ecosystem: There is currently no systematic customer screening across the biotechnology stack—AI model access, gene synthesis ordering, and cloud laboratory use are all effectively ungoverned [19]. Our cross-model bypass finding of 12.1% directly quantifies the credentialing gap for AI access: even if one model has adequate safeguards, the ecosystem does not. Trusted access programs that coordinate refusal policies across model providers would address this gap.

Funders: BioCalibrate satisfies the three filters for fundable interventions [13]: pathogen-agnostic, deployable today, independently useful without universal adoption. It produces rule-in evidence—results unambiguous enough to justify policy action.

6.3 Responsible Evaluation Design

All 338 queries derive from published sources, respecting the "shield of ignorance" [5]. We probe known pathways, not novel ones.

6.4 Limitations

1. **Single-turn evaluation.** Real threat scenarios involve multi-turn conversations where information accumulates. Multi-turn escalation through the attack chain is untested.

2. **Deterministic classification.** Our regex parser prioritizes reproducibility over nuance. It achieves moderate LLM agreement (κ=0.571) with conservative disagreements (see §4.2), but may miss subtle hedging.

3. **Textbook-level queries.** Novel or classified content requires different methodology with appropriate oversight.
4. **No expert panel validation** of pathogen operational risk rankings.
5. **Automated variant generation.** 80.7% validation pass rate indicates reasonable quality, but may not capture full adversarial diversity.
6. **8 models evaluated.** GPT-4o, Gemini 2.5 Pro, and Claude Opus 4.6 excluded due to API access constraints during the evaluation window.

6.5 Future Work

Community-driven model coverage. We are designing a BYOK (bring-your-own-key) contribution pipeline: practitioners would evaluate any model using the existing CLI and submit results for review. Integrity is enforced structurally—raw responses are auditable, scoring is deterministic, and matched pairs prevent selective submission.

Adversarial pathway simulation. Model the acquisition chain as a directed graph—BDL refusal rates as node-level detection probabilities—to identify highest-value intervention combinations.

Cloud lab protocol screening. Extend L3 to structured protocols, evaluating workflows holistically: "is this experiment dangerous?"

Temporal drift tracking. Longitudinal monitoring across model releases.

Multi-turn evaluation. Escalation through the attack chain across conversation turns.

Biological foundation model evaluation. Integrate direct assessment of models like EVO2 and ESM-2, testing whether safety measures persist through fine-tuning [1].

STREAM-compliant reporting. Align with the 28-criteria standard for ChemBio evaluations [18].

Metadata coupling. Explore embedding provenance metadata in bio-AI tool outputs to aid downstream synthesis screening [22].

7 Conclusion

BioCalibrate reveals that current AI biosecurity guardrails are systematically miscalibrated: the best-performing model refuses only 28% of BDL-4 weaponization queries, refusals are driven by pathogen notoriety rather than operational risk (ecosystem FRI = +0.099, $p < 0.05$), and 12.1% of queries can be bypassed by switching models. These findings hold across 8 models from 7 organizations, including three receiving their first published CBRN evaluation, suggesting the problem is rooted in shared safety training approaches rather than individual model failures.

We release BioCalibrate as open infrastructure—a CLI tool, interactive dashboard, and public dataset—to enable continuous monitoring as model capabilities scale. The Digital Biosafety Level framework provides a common language for AI developers, biosecurity governance bodies, and institutional biosafety committees to measure and communicate model safety posture. We hope this work contributes to closing the gap between AI capability growth and biosecurity preparedness.

Code and Data

- **Code repository:** https://github.com/BioCalibrate/BioCalibrate — CLI tool, evaluation pipeline, refusal parser (35+ patterns, 24 tests), scoring metrics, dashboard source, validation study, and reproducibility guide.

- **Dataset:** HuggingFace (lightmate/biocalibrate) — 338 queries with BDL classifications, layer assignments, adversarial-benign pair mappings, pathogen metadata, and LLM-as-judge validation results.
- **Interactive dashboard:** `https://biocalibrate.org` — React SPA with 7 tabs: Overview, Compare, Scorecards, Fear:Risk, Bypass, Methodology, and Contribute.

The tool produces Model Biosafety Scorecards for three primary user groups: AI safety teams evaluating model biosafety before release, policy analysts producing quantitative evidence for biosecurity briefs, and Institutional Biosafety Committees deciding which AI models to approve for laboratory use.

Acknowledgments

Developed during the AIxBio Hackathon (April 24–26, 2026), organized by Apart Research, BlueDot Impact, and Cambridge Biosecurity Hub. AI Biosecurity Tools track sponsored by Fourth Eon Bio. Evaluation layers informed by Alex Gopal (ActiveSight), Gene Olinger (Galveston NL), and James Black (Fourth Eon Bio/Johns Hopkins). Additional insights from Jonas Sandbrink (BioTrust), Connor McGurk (Coefficient Giving), Kevin Esvelt (MIT/SecureBio), Coleman Breen (SecureBio), Oliver Crook (Oxford), Steph Guerra (RAND), and Jamie Yassif (NTI).

References

[1] Black, J., et al. (2025). Adversarial Fine-tuning Rescues Filtered Capabilities in Biological Foundation Models. Fourth Eon Bio.

[2] Breen, C., et al. (2025). BioTier and the SecureBio Evaluation Suite. SecureBio.

[3] Crook, O., et al. (2025). Calibrated Probabilistic Outputs for Biological AI Risk Assessment. Oxford.

[4] CSIS. (2025). Eight Commonsense Actions for Biosafety and Biosecurity.

[5] Esvelt, K. (2025). The Shield of Ignorance: Responsible Evaluation Design for Biosecurity. MIT/SecureBio.

[6] Gopal, A., et al. (2025). Measuring AI-Assisted Uplift in Biology: An RCT. ActiveSight.

[7] Götting, J., et al. (2025). Virology Capabilities Test (VCT). arXiv:2504.16137.

[8] Guerra, S. (2025). AI Tool Chain Risk Assessment for Biological Misuse. RAND.

[9] Guerra, S., et al. (2025). Global Risk Index for AI-enabled Biological Tools. RAND/CLTR.

[10] Knight, C.Q., et al. (2025). FORTRESS: Frontier Risk Evaluation. arXiv:2506.14922.

[11] Li, J., et al. (2025). DarkBench: Benchmarking Dark Patterns in LLMs. arXiv:2503.10728. ICLR Oral.

[12] Li, N., et al. (2024). The WMDP Benchmark. arXiv:2403.03218.

[13] McGurk, C. (2025). Funding Criteria for Biosecurity Technical Interventions. Coefficient Giving.

[14] Mouton, C., et al. (2024). The Operational Risks of AI in Large-Scale Biological Attacks. RAND RR-A2977-2.

[15] Mu, N., Gopal, A., Bradshaw, W. (2025). Recommended Refusal Policies for Highly Capable Models. VCT Appendix A5, SecureBio.

[16] Noever, D. & McKee, F. (2025). Forbidden Science. arXiv:2502.06867.

[17] OpenAI. (2025). GPT-5 System Card. arXiv:2601.03267.

[18] GovAI/SecureBio. (2025). STREAM: ChemBio Evaluation Reporting Standards. arXiv:2508.09853.

[19] Sandbrink, J. (2025). BioTrust: Customer Screening for the Biotech Stack. SentinelBio.

[20] Soice, E.H., et al. (2023). Can LLMs democratize access to dual-use biotechnology? arXiv:2306.03809.

[21] Xie, T., et al. (2024). SORRY-Bench. arXiv:2406.14598.

[22] Yassif, J. (2025). From Technical Innovation to Regulatory Requirement. NTI.

A Limitations and Dual-Use Considerations

Limitations

False positives/negatives. The deterministic regex parser (κ=0.571 vs. LLM judge) may misclassify nuanced responses. It errs conservative—under-counting refusals rather than over-counting—so reported compliance rates are upper bounds. However, some models embed safety caveats ("consult a biosafety officer") within otherwise compliant responses; our parser may classify these as safe completions when human judgment would rate them differently.

Edge cases. The 338-query dataset covers the acquire-cultivate-disseminate chain but cannot exhaustively represent all adversarial phrasings. The 80.7% automated validation rate indicates reasonable but imperfect quality control. Multi-turn conversations, system prompt manipulation, and few-shot jailbreaking techniques are untested.

Scalability. Evaluating additional models requires API access and compute. The BYOK contribution pipeline is designed but not yet deployed; scaling to continuous monitoring depends on community adoption.

Dual-Use Risks

BioCalibrate's query dataset could theoretically be repurposed to identify which models are most permissive for dangerous queries. We mitigate this through design: (1) all queries derive from published sources already available in the biosecurity literature—the dataset contains no novel attack methods or information not already public; (2) the tool's primary output is a *risk surface* identifying weaknesses, not a guide for exploiting them; (3) the matched adversarial-benign pair design means the dataset is most useful for measuring calibration quality, not for extracting dangerous content.

The cross-model bypass analysis identifies specific model combinations that create vulnerabilities. We report the aggregate bypass rate (12.1%) rather than publishing per-query, per-model bypass matrices, to provide actionable evidence for ecosystem-level interventions without creating a routing guide for adversaries.

Responsible Disclosure

All findings have been reported in aggregate. We do not publish specific prompts that reliably extract dangerous content from specific models. The 338-query dataset is released publicly because it contains only published-source queries; the value of the dataset for improving safety systems outweighs the marginal risk given that all underlying information is already public.

We recommend that AI developers use BioCalibrate's Model Biosafety Scorecards to identify specific BDL-level failure categories in their models and prioritize BDL-4 compliance improvements. We further recommend that ecosystem-level coordination bodies (e.g., Frontier Model Forum) use the cross-model bypass analysis to inform shared refusal policy development.

Ethical Considerations

This work evaluates AI model safety systems without generating novel biological threat information. The "shield of ignorance" [5] is respected throughout: we probe models on *known, documented* pathways from published literature [14, 20], never generating or publishing novel weaponization methods. Operational risk scores (e.g., influenza 0.90, Ebola 0.20) are derived from published biosecurity frameworks, not independent risk assessment.

The evaluation was conducted without jailbreaking, prompt injection, or adversarial prompt engineering—testing baseline model behavior as a user would encounter it. This design choice means our results represent a lower bound on actual vulnerability; a motivated adversary using jailbreaking techniques would likely achieve higher compliance rates.

Suggestions for Future Improvements

Expert panel validation of operational risk rankings would strengthen the BDL framework's authority. Expanding to multi-turn evaluation, system prompt manipulation, and jailbreaking resistance would provide a more complete safety picture. Integration with SecureBio's BioTier and STREAM reporting standards [18] would improve interoperability with the broader biosecurity evaluation ecosystem. A responsible disclosure process for newly discovered model-specific vulnerabilities should be formalized before expanding the evaluation scope.

LLM Usage Statement

LLMs were used in two capacities during this project: (1) **dataset generation**—DeepSeek-V3 generated query variants from hand-crafted seeds, and Qwen3-235B validated variant quality (80.7% pass rate); (2) **validation**—Claude Sonnet 4 served as an independent LLM judge for the refusal classification validation study (κ=0.571, n=160). LLM assistance was also used for brainstorming approaches, drafting sections, and debugging code. All quantitative results, claims, and analysis were independently verified. The final paper was primarily written and reviewed by the author.