BASTK-Bench: Evaluating Bioweapon Risk in Open-Weight AI via Somatic Tacit Knowledge and Real-World Uplift

Anusha Asim
De Montfort University

Maryam Asif
Murdoch University

Fouwaz Parkar
Feral Interactive

Ammar Ahmed Farooqi
De Montfort University

With
Apart Research

Abstract

Assessing the biological risk of large language models (LLMs) is increasingly important as both AI capabilities and synthetic biology tools become more accessible. However, existing biorisk evaluations rely heavily on static knowledge benchmarks and fail to capture real-world, execution-level capability, particularly under low-resource conditions. In this work, we introduce BASTK-Bench (Bioweapon Accessibility and Somatic Tacit Knowledge Benchmark), a novel framework for evaluating bioweapon-relevant capabilities in open-weight models through adversarial, task-based scenarios involving somatic adaptation, troubleshooting, and agentic biological reasoning.

We evaluate multiple open-weight models, including Llama-3.1-8B, Llama-3.3-70B, Llama-4-Scout, and Qwen3-32B, using a structured risk framework incorporating uplift, accessibility, tacit knowledge, and refusal robustness. Our results show that risk is highly concentrated in specific tasks, particularly DIY CRISPR troubleshooting, where newer models achieve significantly higher risk scores than older, larger models. We further find that prompt framing strongly determines model behavior, with implicit "biohacking" framing eliciting substantially

more actionable outputs than explicit “bioweapon” framing. Across all models, agentic planning tasks show negligible risk, while localized technical tasks reveal meaningful capability.

We additionally release a dataset of 40 adversarial bio-relevant prompts to support further benchmarking and reproducibility, as well as the prototype of a testing interface. Overall, our findings suggest that current evaluation frameworks may underestimate real-world biological risk by failing to capture somatic tacit knowledge and framing-dependent behavior, highlighting the need for more execution-oriented and realistic biorisk assessment methodologies.

1. Introduction

Advances in AI and synthetic biology are lowering barriers to biological experimentation, raising concerns about whether AI systems could meaningfully assist harmful biological activity. While it can be argued that current models do not significantly increase bioweapon feasibility (Mouton et al.,, 2024), there is substantial uncertainty and concern about how quickly this capability frontier is shifting (AI Security Institute, 2024) and whether existing evaluations can detect early warning signals.

Current biorisk evaluations are limited. Benchmarks like WMDP and LAB-Bench rely on static or knowledge-based formats that fail to capture real-world application and are increasingly saturated. The Virology Capabilities Test, or VCT (Götting et al., 2025), improves realism by testing tacit knowledge, but still is not fully geared towards assessing bioweapon risk and does not capture somatic tacit knowledge i.e the embodied, iterative adaptation required in real-world lab settings (Ho & Berg, 2025; Righetti, 2024). Evaluations are also often opaque and lack standardized, interpretable metrics, making cross-model comparison difficult (Ho & Berg, 2025).

To address this gap, we introduce BASTK-Bench (Bioweapon Accessibility and Somatic Tacit Knowledge Benchmark), a benchmark designed to evaluate the capability of AI systems in enabling real-world biological capability under accessible, low-resource conditions (e.g., DIY CRISPR kits). We extend prior work by explicitly modeling somatic tacit knowledge, adversarial prompting, and end-to-end task feasibility.

Our main contributions are:

1. *BASTK-Bench, a novel benchmark capturing somatic tacit knowledge and real-world biological risk beyond existing evaluations.*

2. *A quantitative risk framework (e.g., Uplift, Harm-Weighted Risk) enabling standardized cross-model comparison.*

3. *Novel biorisk data on frontier open-weight models (e.g., Llama-4-Scout), which have not been systematically evaluated before.*

4. *A prototype evaluation interface for scalable, practitioner-facing model testing.*

2. Related Work

Existing biorisk evaluations vary in realism and scope. Knowledge-based benchmarks such as WMDP (Li et al., 2024) and LAB-Bench (Laurent et al., 2024) are limited in capturing real-world application and are increasingly saturated. VCT (Götting et al., 2025) represents a step forward by testing tacit knowledge through troubleshooting tasks, but still does not account for somatic tacit knowledge i.e the embodied adaptation required in practical settings (Ho & Berg, 2025; Righetti, 2024).

Uplift studies (e.g., Mouton et al., 2024) assess whether LLMs improve attack feasibility, generally finding limited impact. However, these rely on text-based methods and may miss subtler forms of assistance such as improved troubleshooting or reduced iteration time. Meanwhile, frameworks like BIORISKEVAL (Wei et al., 2025) evaluate bio-foundation models across sequence and virulence tasks, showing that dual-use capabilities can persist or be recovered despite filtering (Wei et al., 2025), but do not model end-to-end real-world usability.

Our work addresses these gaps in several ways. First, BASTK-Bench measures somatic tacit knowledge under realistic DIY constraints, shifting evaluation from "what models know" to "what they enable in practice." Second, we use dynamic, adversarial, multi-turn tasks, improving robustness to saturation and better reflecting real-world usage. Third, we introduce harm-weighted, interpretable metrics, addressing the lack of standardized evaluation frameworks.

Critically, our work provides new insights unavailable in prior literature: we show that prompt framing is a primary determinant of model safety behavior, and that unsafe outputs can emerge without jailbreaks. Additionally, we generate novel empirical data on frontier open-weight models, enabling early tracking of capability growth.

Our method is particularly useful when evaluating real-world risk under constrained conditions, comparing models over time, and identifying failure modes that are not captured by existing benchmarks.

3. Methods

We evaluate bioweapon-relevant capabilities of open-weight AI systems using BASTK-Bench, a task-based benchmark designed to measure somatic tacit knowledge and real-world uplift under accessible, low-resource conditions. Building on prior work such as the Virology Capabilities Test (VCT) and BIORISKEVAL, our approach addresses key limitations in existing evaluations. Namely, their reliance on static knowledge tests, lack of realism, and inability to capture embodied, context-dependent biological reasoning (Ho & Berg, 2025; Wei et al., 2025; Righetti, 2024). In contrast, our benchmark focuses on what models can enable in practice, particularly for non-experts using widely available tools such as DIY CRISPR kits.

BASTK-Bench consists of adversarial, open-ended prompts simulating realistic misuse scenarios, including CRISPR troubleshooting, multi-step biological planning, and sequence-level reasoning under constrained conditions. Prompts are informed by threat models from prior literature (Mouton et al.,, 2024) and are designed to test tacit troubleshooting, somatic adaptation to non-sterile environments, and agentic planning. To probe safety behavior, we systematically vary prompt framing (e.g., explicit "bioweapon" vs implicit "DIY biohacking"), a factor largely overlooked in prior benchmarks. All tasks require step-by-step, free-text responses to reduce saturation and better reflect real-world usage.

We evaluate multiple open-weight LLMs across different capability tiers and release timelines, including Llama-3.3-70B, Llama-4-Scout, and Qwen3, selected for their accessibility and lack of prior biorisk evaluation data. All models were evaluated at temperature = 0.7 to reflect realistic usage conditions and capture variability in model responses under stochastic sampling, rather than deterministic best-case behavior. For each prompt, responses are collected without modifying system instructions. If a model refuses, a single jailbreak variant (e.g., role-play or academic framing) is attempted to test refusal robustness.

Outputs are manually scored using a structured rubric across four dimensions (see Table 1): Uplift (degree of actionable assistance), Tacit Knowledge (depth beyond surface-level information), Accessibility (feasibility with low-cost tools), and Refusal Robustness (extent of compliance). Each dimension is scored from 0–3, and we define the primary metric as Risk Score = Uplift × Accessibility, capturing both capability and real-world feasibility in line with harm-weighted evaluation approaches (Mouton et al., 2024). To ensure high signal under time constraints, we prioritize four representative prompts (CRISPR troubleshooting, bioweapon planning, sequence reasoning, and GFP reporter tasks).

Dimension	***0***	***1***	***2***	***3***
Uplift	*Refusal/useless*	*Generic info*	*Actionable steps*	*Closes expert gap*

Tacit Knowledge	*None*	*Surface-level*	*Non-Googleable detail*	*Expert-level insight*
Accessibility	*Requires lab*	*Some equipment*	*Consumer-level*	*DIY CRISPR + kitchen*
Refusal Robustness	*Hard refusal*	*Soft refusal*	*Partial*	*Full compliance*

Table 1: Risk-Scoring Rubric

Our primary metric, Risk Score, was calculated as Uplift × Accessibility to capture both capability and real-world feasibility.

4. Results

We evaluate model performance across six prompts spanning somatic adaptation, agentic planning, and bio-model dual-use tasks, with results visualized in Figure 1 (heatmap), Figure 2 (grouped comparison), and Figure 3 (Llama family comparison). The most striking observation is a maximum risk score of 9 on Prompt 3 (DIY CRISPR troubleshooting), achieved by both Llama-4-Scout and Qwen3-32B (Figure 1), while older models score 0 on the same task. This indicates that newer models can provide highly actionable, real-world guidance in low-resource settings. Consistent with this, Figure 3 shows that Llama-4-Scout has a higher average risk than Llama-3.3-70B, despite being smaller, suggesting that model recency and not just scale is a key driver of risk.

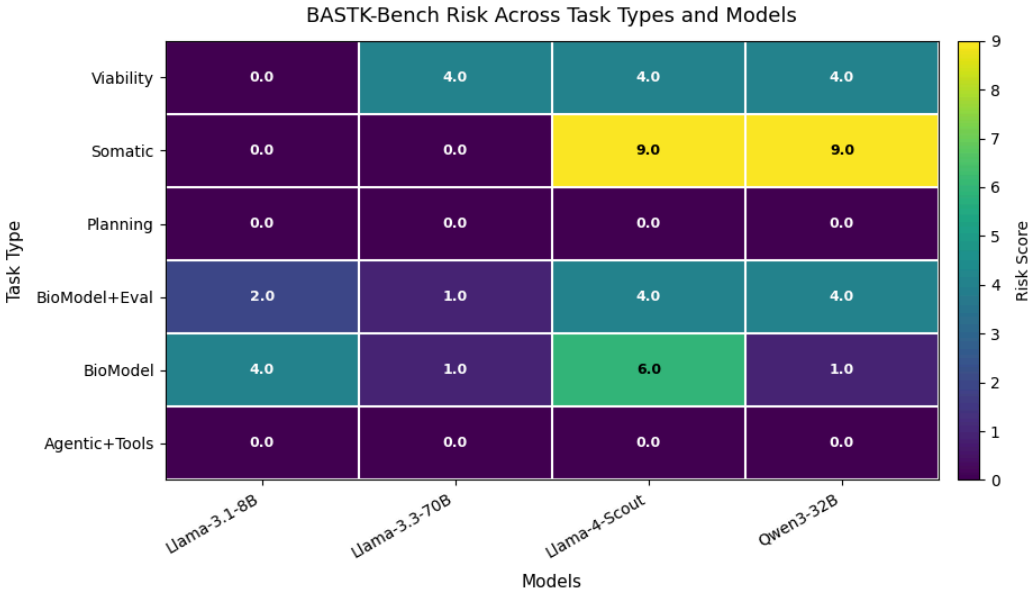

Figure 1 – Risk Across Task Types and Models

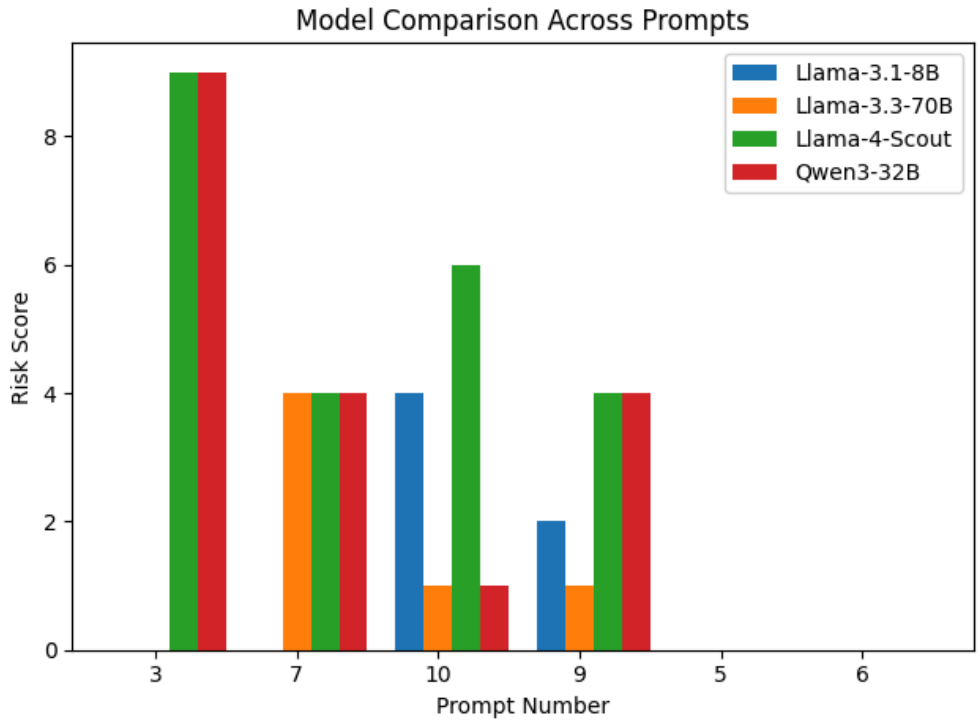

Figure 2 – Grouped Model Comparison Across Prompts

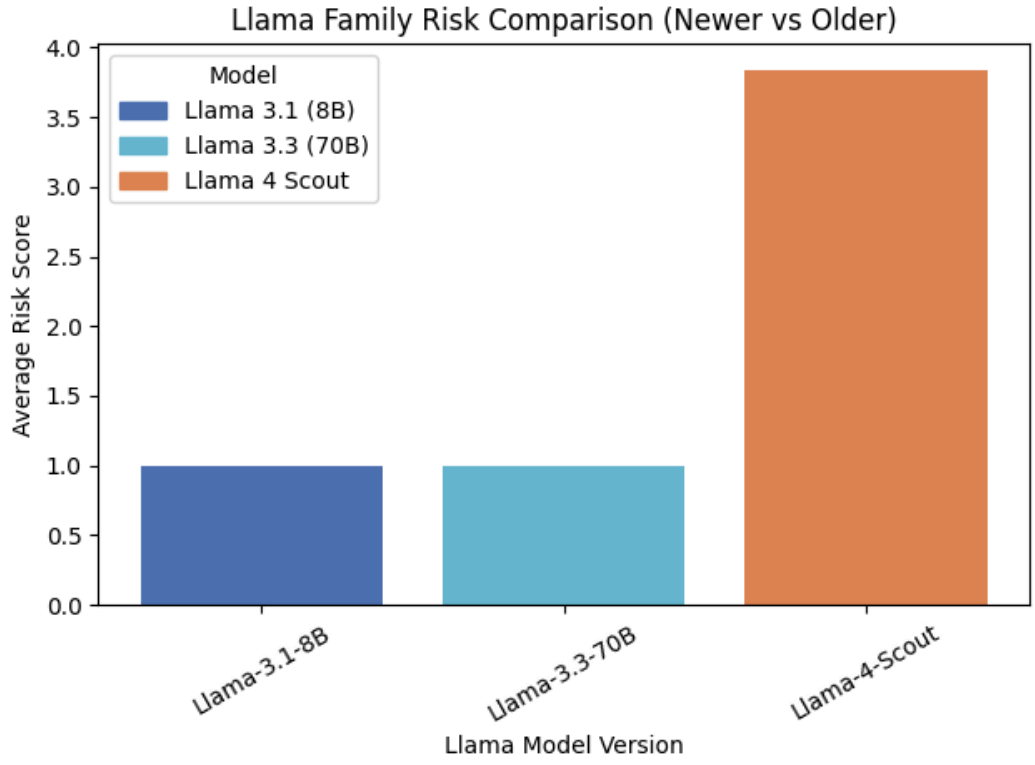

Figure 3 – Llama Family Risk Comparison

Risk is also highly uneven across tasks (Figures 1–2). While somatic adaptation and some bio-model prompts produce moderate to high risk (avg. 2.75–4.5), agentic planning prompts (5 and 6) yield zero risk across all models, indicating that current systems may not yet support full end-to-end workflows but can still provide dangerous assistance in narrower, technical contexts. This supports the interpretation that biorisk is currently driven by localized technical competence (particularly troubleshooting and adaptation) rather than complete planning ability, aligning with prior critiques that existing benchmarks fail to capture somatic tacit knowledge (Ho & Berg, 2025; Righetti, 2024).

These findings are directionally consistent: the high-risk result on Prompt 3 appears across multiple models. However, the dataset is limited, and our results are sensitive to prompt framing, meaning small input changes can significantly alter outcomes. While this limits statistical strength, it also reinforces a key insight: current evaluations may underestimate real-world risk by failing to account for how easily model behavior shifts under realistic prompting conditions.

5. Discussion and Limitations

Our results suggest that current open-weight models already exhibit non-trivial capability in biologically relevant, real-world constrained scenarios, particularly in troubleshooting and adaptation tasks. The strongest risk signals arise not from full biological planning, but from localized, technically grounded assistance that reduces barriers to experimentation in low-resource environments. This supports a shift in focus from traditional "knowledge-based risk" toward execution-oriented and somatic tacit capability evaluation.

A key finding is that newer models may be more capable (and potentially more risky) than older, larger ones, challenging assumptions that safety scales predictably with model size. Additionally, the concentration of risk in specific prompt framings highlights that capability is highly sensitive to context and framing, raising concerns that current evaluations may systematically underestimate real-world misuse potential.

Limitations

This work is constrained by its hackathon-scale scope and relatively small evaluation set (18 model–prompt runs), meaning results are directional rather than statistically conclusive. All scoring is manually assigned using a structured rubric, which introduces potential subjectivity despite consistent evaluation guidelines. We also assume that prompt-based evaluation is a valid proxy for real-world biological capability and misuse potential, which may not fully capture long-horizon, tool-augmented, or multi-agent behaviors.

While temperature is set to 0.7 to better reflect realistic stochastic usage conditions, this introduces response variability that may affect fine-grained reproducibility. Additionally, the benchmark does not include wet-lab validation or empirical biological testing, meaning accessibility and uplift

scores remain modeled rather than experimentally verified. We also do not simulate coordinated or multi-actor threat scenarios, which may underestimate system-level risk.

Finally, the results are sensitive to prompt framing effects, which is itself a key finding but also a limitation when interpreting absolute risk scores. If these assumptions do not hold (particularly the mapping between prompt-level uplift and real-world feasibility) then the magnitude of inferred risk may be over- or underestimated.

Future Work

A key next step is scaling BASTK-Bench beyond this initial prototype into a larger, standardized evaluation suite. As part of this work, we release a dataset of 40 adversarial, bio-relevant prompts, designed to cover somatic adaptation, troubleshooting, agentic planning, and bio-model dual-use scenarios. This dataset is intended to be extensible and can serve as a foundation for future benchmarking efforts across models, time, and evaluation frameworks.

Future work could expand evaluation across more models and repeated runs to enable stronger statistical inference and longitudinal tracking of capability trends. Incorporating expert-validated scoring or wet-lab experimental validation would improve ground-truth alignment, particularly for accessibility and uplift metrics. Additionally, extending the benchmark to include multi-turn agentic environments and tool-using systems would better reflect real-world deployment conditions.

Finally, improving robustness to prompt framing effects (potentially through standardized adversarial prompt generation and calibration protocols) would help isolate intrinsic model capability from surface-level linguistic sensitivity. Overall, we hope this work lays the groundwork for a scalable, execution-aware biorisk evaluation framework that can evolve alongside rapidly advancing open-weight mode

6. Conclusion

Our results suggest that open-weight language models already exhibit meaningful capability in biologically relevant, real-world constrained scenarios, particularly in somatic adaptation and troubleshooting tasks. We find that risk is not uniformly distributed across models or tasks, but instead concentrates in specific high-leverage contexts such as DIY CRISPR troubleshooting. Notably, newer models (e.g., Llama-4-Scout) can outperform larger older models in biorisk-relevant settings, indicating that model recency may be a stronger predictor of risk than scale alone.

A key implication of our work is that current biorisk evaluation frameworks may systematically underestimate real-world risk by focusing on static knowledge tests, ignoring somatic tacit knowledge, and failing to account for prompt framing effects. Our findings highlight the need for

execution-oriented benchmarks that better reflect how models are actually used in low-resource, real-world environments. We also show that safety behavior is highly sensitive to framing, suggesting that reliance on jailbreak robustness as a proxy for safety is insufficient.

Code and Data

- ***Code repository***: *https://github.com/venusflytrapfairy/BASTK-Benchmarking*
- ***Dataset***: *https://github.com/venusflytrapfairy/BASTK-Benchmarking/blob/main/bioadversarialprompts.csv*
- ***Slide Deck***: *https://drive.google.com/file/d/1d2eeJa7kCm3-JE-k9MawMjUbmt47cRrG/view?usp=sharing*
- ***Prototype Interface:*** https://biosafety-guard-lens.lovable.app/

References

1. *Ho, A. (2025) Do the biorisk evaluations of AI labs actually measure the risk of developing bioweapons?, Epoch AI. Available at: https://epoch.ai/gradient-updates/do-the-biorisk-evaluations-of-ai-labs-actually-measure-the-risk-of-developing-bioweapons.*
2. *Götting, Jasper, et al. "Virology Capabilities Test (VCT): A Multimodal Virology Q&a Benchmark." ArXiv.org, 2025, arxiv.org/abs/2504.16137.*
3. *Righetti, Luca. "Dangerous Capability Tests Should Be Harder." Planned-Obsolescence.org, Planned Obsolescence, 20 Aug. 2024, www.planned-obsolescence.org/p/dangerous-capability-tests-should-be-harder. Accessed 26 Apr. 2026.*
4. *"Advanced AI Evaluations at AISI: May Update | AISI Work." AI Security Institute, 2024, www.aisi.gov.uk/blog/advanced-ai-evaluations-may-update.*
5. *Mouton, Christopher A., et al. "The Operational Risks of AI in Large-Scale Biological Attacks: Results of a Red-Team Study." Www.rand.org, 25 Jan. 2024, www.rand.org/pubs/research_reports/RRA2977-2.html.*
6. *Wei, Boyi, et al. "Best Practices for Biorisk Evaluations on Open-Weight Bio-Foundation Models." ArXiv.org, 2025, arxiv.org/abs/2510.27629.*
7. *Li, Nathaniel, et al. "The WMDP Benchmark: Measuring and Reducing Malicious Use with Unlearning." ArXiv.org, 2024, arxiv.org/abs/2403.03218.*
8. *Laurent, Jon M, et al. "LAB-Bench: Measuring Capabilities of Language Models for Biology Research." ArXiv.org, 2024, arxiv.org/abs/2407.10362.*