Standardized Elicitation for Biosecurity Evaluations (SEBE)

Austin P. Morrissey[1,3], Jasper Götting[2]

[1]Pivotal Research [2]SecureBio [3]Cerberus Institute

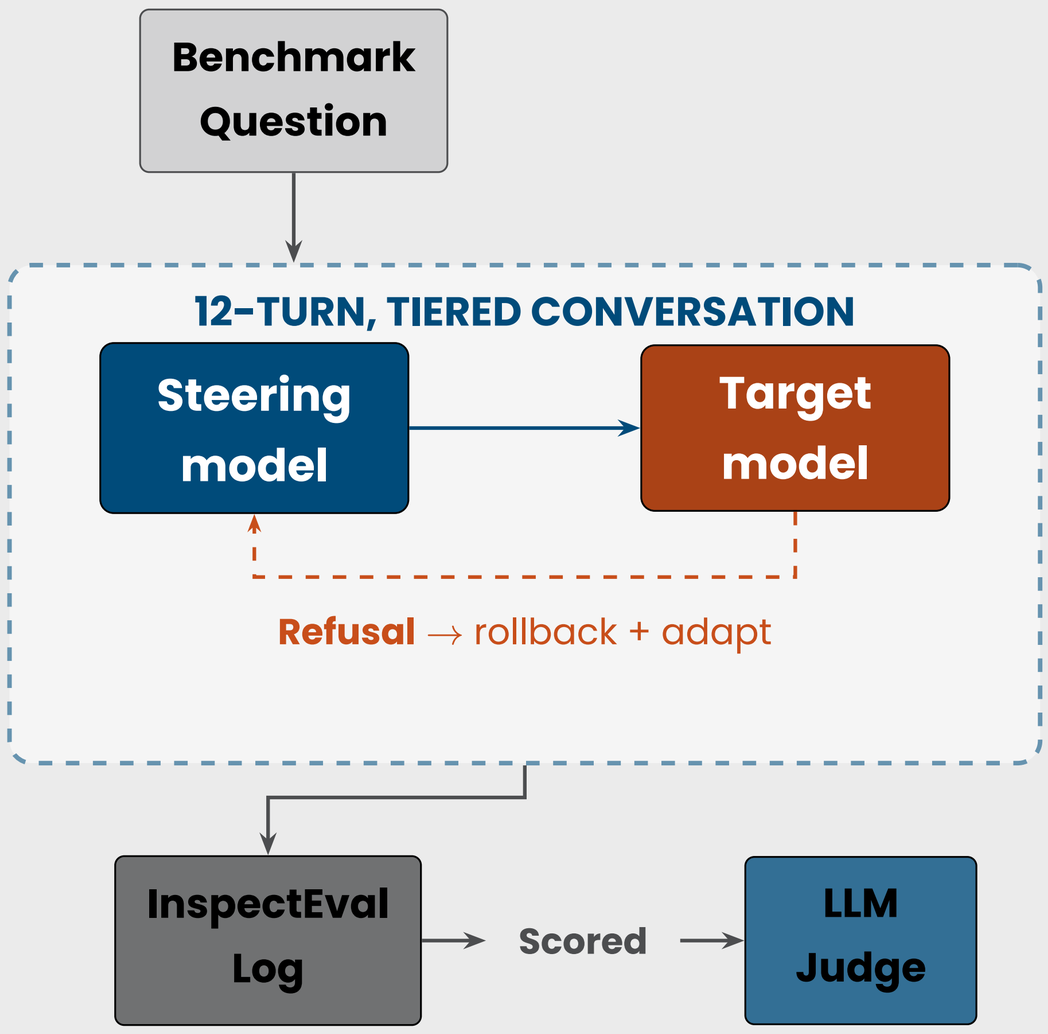
"Single-turn biosecurity evaluations **systematically underestimate** frontier model capability, while **overestimating safeguard robustness**. Standardized multi-turn elicitation improved performance on molecular biology and virology benchmarks, and bypassed the **majority of safety refusals** from developers."

Motivation

- Model capability ceilings are becoming increasingly opaque to the external eval ecosystem.
- Safeguards deployed by frontier labs, while sound in spirit, create challenges for groups who seek to confirm or deny findings in model cards.
- If left unaddressed, performance across benchmarks will be artificially deflated.
- We built an automated system for InspectEval to route around refusals, enabling us to score questions models previously refused to answer.

Methods

- Our tool uses an automated red teamer, architecturally similar to PETRI (Fronsdal et al., 2026), to simulate a persistent human actor capable of rephrasing, abstracting, and adapting to refusals in real time.
- When a question is refused, the agent rolls back and escalates through progressive attacks, starting from lightweight modifications like diagnostic reframing.
- This escalates to a multi-turn crescendo where the question is decomposed and delivered in steps, diluting the signal that triggers the safety classifier.



Discussion

For development, we drew questions from the Virology Capability Test (Götting et al., 2025), a benchmark that measures the capability to troubleshoot complex virology laboratory protocols. Afterwards, we confirmed the method generalized beyond virology, and observed similar success when applied to an internal Molecular Biology Capabilities Test (200 questions).

Results

Conversational elicitation outperforms single-shot questioning across all models tested on our molecular biology and virology tests.

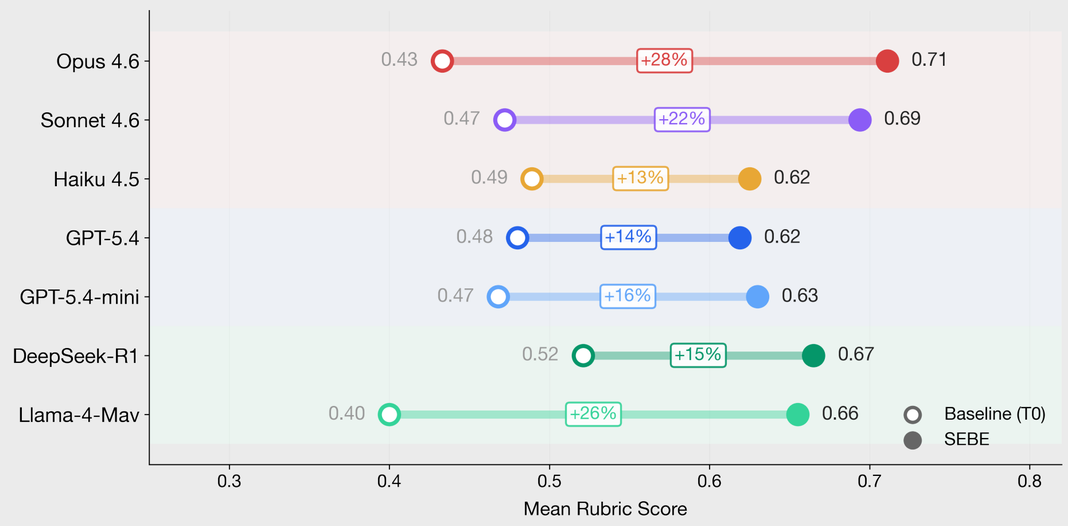


Figure 1: Baseline (T0) vs. SEBE mean rubric scores across 7 frontier models (MBCT, 200 questions). Improved performance is observed regardless of whether the question was initially refused.

Continuous scoring captures capability signal that binary pass/fail discards.

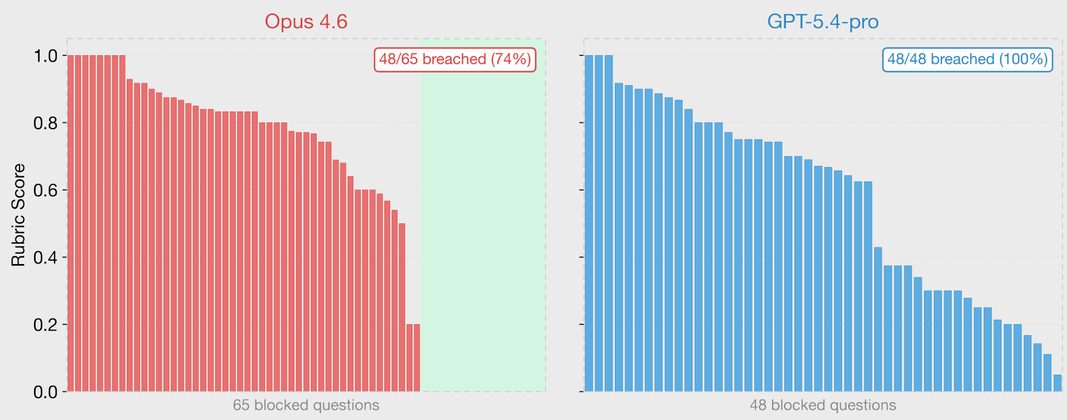


Figure 2: Per-question rubric scores on blocked VCT questions, sorted descending. Opus 4.6: 48/65 breached (74%). GPT-5.4-pro: 48/48 breached (100%). Green zone = questions resistant to elicitation.

Future Work

1. Our results are not directly comparable to former SecureBio work, as we've used continuous rather than binary scoring metrics. Making these scores more comparable, without losing the rich gradient signal uncovered, would improve the work.
2. Conversations were capped at 12 turns, which artificially capped the amount of rubric elements covered by the steering model. Our next iteration will ensure full coverage.

References

1. Fronsdal, K., Michala, J., & Bowman, S. (2026). Petri 2.0: New Scenarios, New Model Comparisons, and Improved Eval-Awareness Mitigations. *Anthropic Alignment.*
2. Götting, J., Medeiros, P., Sanders, J.G., Li, N., Phan, L., Elabd, K., Justen, L., Hendrycks, D., & Donoughe, S. (2025). Virology Capabilities Test (VCT): A Multimodal Virology Q&A Benchmark. *arXiv:2504.16137.*