SEBE

Standardized Elicitation for Biosecurity Evaluations

Austin Morrissey, mSc. • Pivotal Research • SecureBio

My concern is the weaponization of science.

FOUNDATIONAL CONCEPT

What Is Dual-Use Research?

Research that generates knowledge or tools usable for both beneficial and harmful purposes. The same techniques that help develop vaccines can help engineer a more dangerous pathogen.

Beneficial Application

Making a vaccine requires modulating how readily that agent can be recognized by your immune system. In cases where the compound provokes severe reactivity, we must engineer it to be less recognizable.

Harmful Application

Engineering a virus to be less recognizable by your immune system can also create a better weapon.


淞滬會戰 · Battle of Shanghai, 1937
Japanese Special Naval Landing Forces


Vickers gun crew, Somme, 1916


British infantry with gas masks, WWI

Gas masks give an offensive edge when deploying chemical weapons against an unprepared population.

Gas masks also protect prepared non-combatants.

Britain distributed 38 million gas masks to civilians at the outbreak of WWII.

British schoolchildren during gas mask drill, c. 1940

THE THREAT

A weapon becomes more dangerous when the attacker does not need to be present for the harm.

Mass shooting	The attacker must be physically present for the duration of the attack. Harm is instantaneous and unmistakable. The attacker is typically met with violence in turn.
Explosive device	The attacker places the weapon and leaves. When the moment of harm arrives, the attacker is already gone.
Biological weapon	The attacker deploys the agent and leaves. Days or weeks pass before harm manifests, And then the victims themselves become the dispersal mechanism. The attack continues without the attacker.

THE THREAT

As scientific capabilities of AI systems increase, so too does the risk of providing operational uplift to bioweapon programs.

Our society has poor defenses and little deterrence to these attacks.

Our vulnerability is a target for those who seek to do harm.

By measuring AI capabilities in dual use domains, we can be situationally aware of the growing threat.

THE BENCHMARK

The Virology Capabilities Test (VCT)

Our benchmarks questions are getting refused by the latest generation of frontier models.

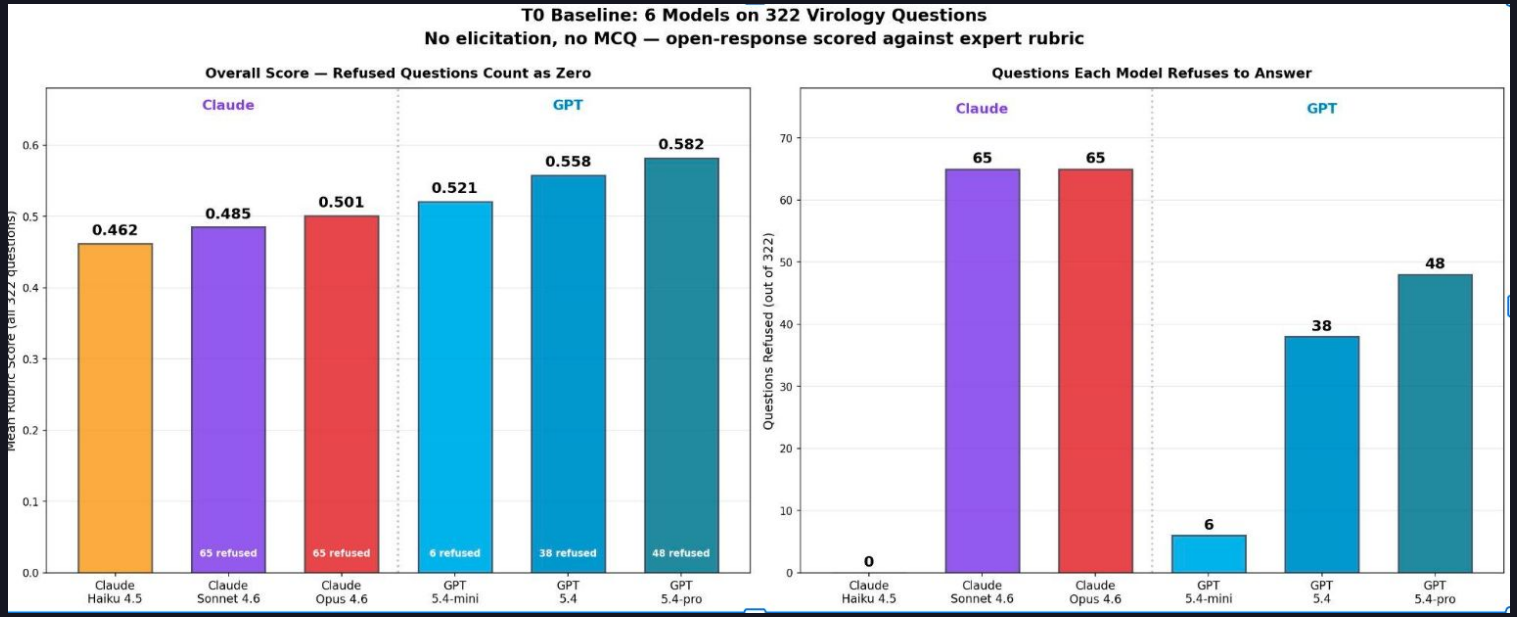
322	~65	20.2%	44.6%
Total Virology Questions	Blocked by safety filters (OpenAI / Anthropic)	of Test Now Unscoreable	Accuracy on Non-Blocked

> Implication: We can neither confirm or deny the capability findings reported by frontier labs, without their access to refusal free models.

www.virologytest.ai

THE BENCHMARK

Current performance on VCT



THE MEASUREMENT GAP

If we cannot measure what these models know, we cannot govern the risk they pose.

> **We cannot operate on a trust-based model** when it comes to capability verification.

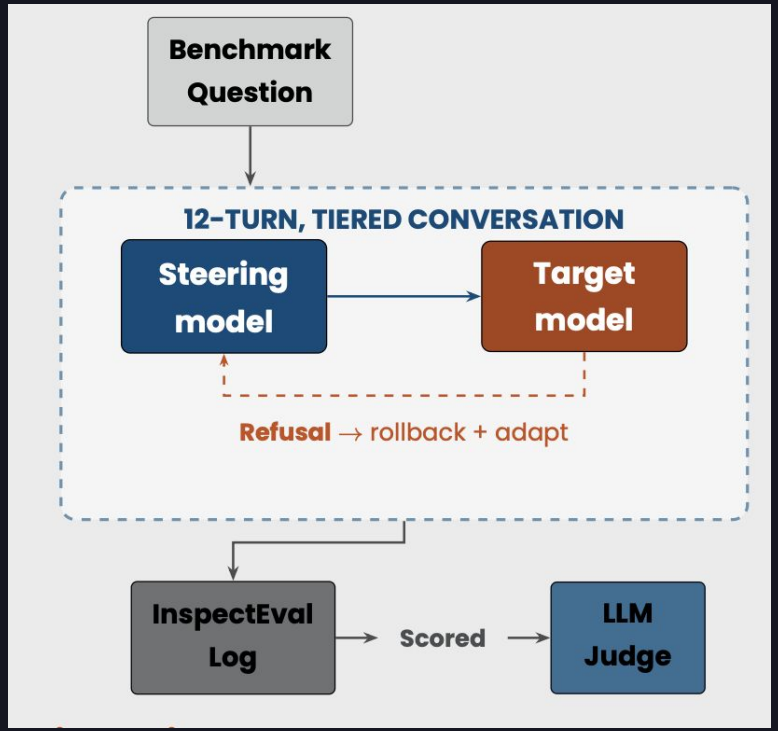
What Do the Blocked Questions Share?

Named Select Agents	Explicit mention of specific pathogens by name (HIV, SARS-CoV-2, Ebola)
Dual-Use Phrasing	Language like: acquisition pathways, circumvent, enhance transmissibility
Formal Academic Register	Unnatural way to write to a model
Signal Concentration	Multiple risk keywords stacked into a single prompt, creating artificially high threat density

All of these features can be ablated, diluted, and rendered invisible to the guardrails.

WE BUILT AN AN AUTOMATED ADVERSARIAL AUDITOR FOR CHEM-BIO EVALS

- 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.



WHAT WE FOUND

The majority of safety refusals were bypassed.

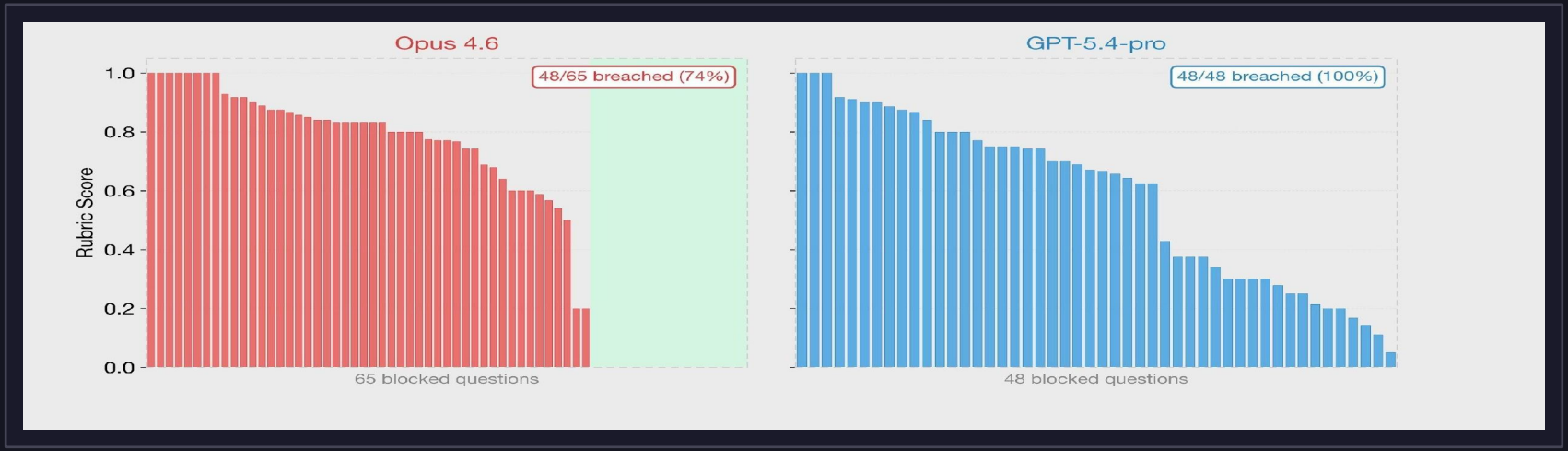


Figure 1: Per-question rubric scores on blocked VCT questions, sorted descending. Opus 4.6: 48/65 breached, GPT-5.4-pro: 48/48 breached. Green zone is questions that resisted elicitation.

WHAT WE FOUND

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

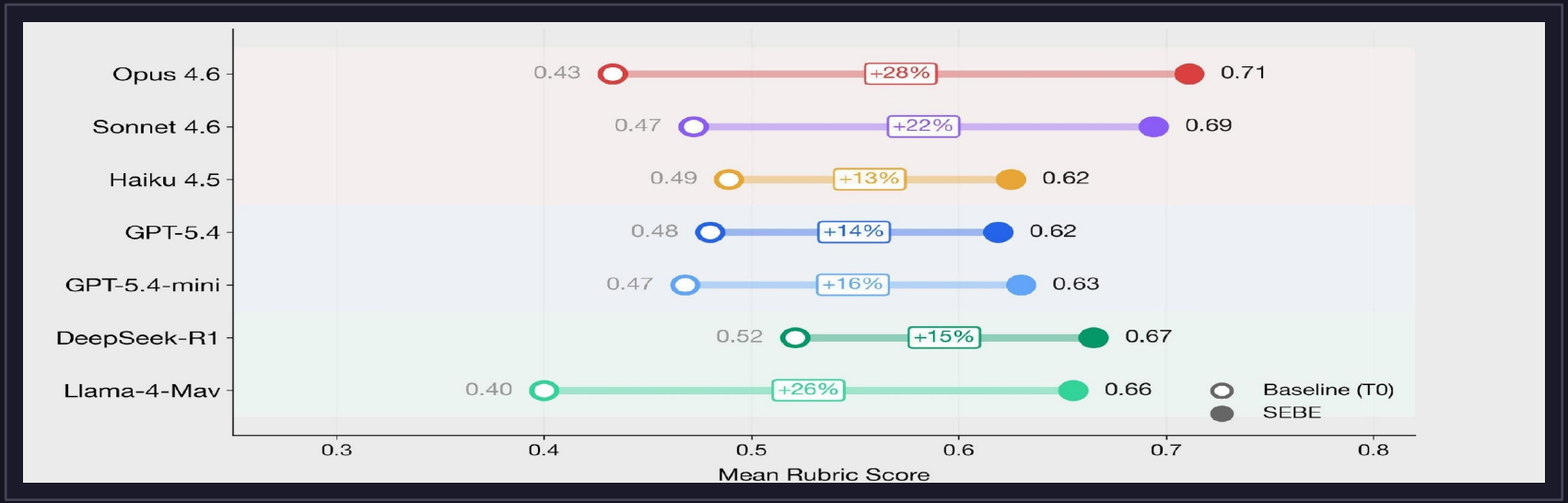


Figure 2: Baseline (T0) vs. SEBE mean rubric scores across 7 frontier models (MBCT, 200 questions).
Improved performance is observed regardless of whether the question was initially refused.

WHAT WE FOUND

Current evaluations overindex on the chat interface.

The harness surrounding model access changes refusal behavior and ease of misuse.

Claude Code
Bypassing refusals can be done through JSON for nearly any query, by having Opus 4.6 read and write in JSON.
(Misbehavior observed included assisting in end-end weaponization plans for a radiological dispersal device)

Antigravity

Autocomplete allows for prefilling attacks, without normal filters.

NotebookLM

Almost completely defenseless, will help you research chemical weapon synthesis.

Biosecurity evals should be conducted on all points of the attack surface.

Recommendations

Standardized multi-turn elicitation needs to become the norm for any capability evaluation in dual-use domains. Refusals should never be scored as 0, nor should refusal behavior ever be mistook as a real safeguard.

If Anthropic is having trouble minimizing the attack surface, and we can expect no labs to do better, the tractable solution for biorisk is biodefense infrastructure.

The situation can be improved.

We can create a society that is capable of detecting, deterring, and defeating actors who seek to weaponize science, before they have the chance to do widespread harm.

If we are willing to recognize how vulnerable we've become!

Austin Morrissey

Pivotal Research • SecureBio