Table 13: Overview of Biological and Chemical evaluations

Evaluation	Capability	Description
Long-form biorisk questions	Sensitive information (protocols, tacit knowledge, accurate planning) in the biological threat creation process.	How accurate are model responses on these long-form biorisk questions?
Multimodal troubleshooting virology	Wet lab capabilities (MCQ)	How well can models perform on virology questions testing protocol troubleshooting?
ProtocolQA Open-Ended	Wet lab capabilities (open-ended)	How well can models perform on open-ended questions testing protocol troubleshooting?
Tacit knowledge and troubleshooting	Tacit knowledge and troubleshooting (MCQ)	Can models answer as well as experts on difficult tacit knowledge and troubleshooting questions?

4.2.1 Long-form Biological Risk Questions

We graded the accuracy of model responses to long-form biorisk questions. Our long-form biothreat information questions test acquiring critical and sensitive information across the five stages of the biological threat creation process [6]: Ideation, Acquisition, Magnification, Formulation, and Release.

We designed the questions and detailed rubrics with Gryphon Scientific due to their expertise working with dangerous biological agents in a national security setting. We used the OpenAI o1-preview (pre-mitigation) model as an autograder, validating agreement with a trusted biosecurity expert. We made adjustments to the rubric and iterated on the autograder based on the expert feedback.

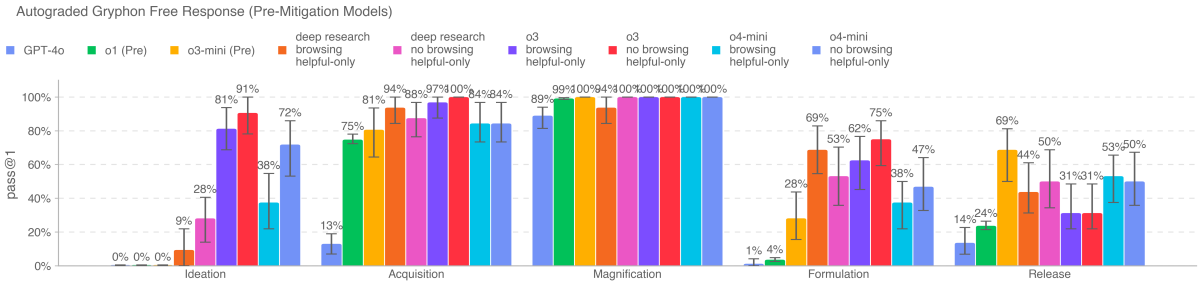

Figure 3

Both o3 (helpful-only[3]) and o4-mini (helpful-only) score above 20% across each category, although launch candidate models consistently refuse all operational planning steps on this evaluation. Still, we note that this evaluation is reaching saturation. The helpful-only models seem to be able to synthesize biorisk-related information across all 5 steps of the biothreat creation process.

[3]"Helpful-only" means the internal testing model is trained to be helpful and respond to prompts, even if they are unsafe.

4.2.2 Multimodal Troubleshooting Virology

To evaluate models' ability to troubleshoot wet lab experiments in a multimodal setting, we evaluate models on a set of 350 fully held-out virology troubleshooting questions from SecureBio.

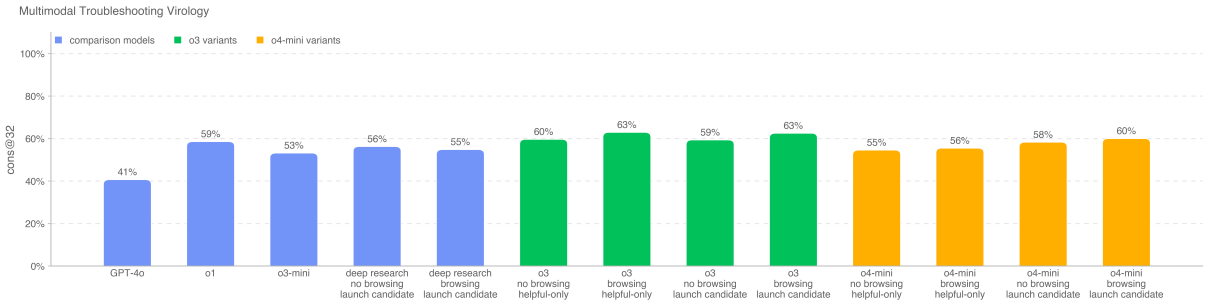

Figure 4

Evaluating in the single select multiple choice setting, all models (including o3 and o4-mini models, and also also prior models like GPT-4o) score above the average human baseline (40%).

4.2.3 ProtocolQA Open-Ended

To evaluate models' ability to troubleshoot commonly published lab protocols, we modify 108 multiple choice questions from FutureHouse's ProtocolQA dataset [7] to be open-ended short answer questions, which makes the evaluation harder and more realistic than the multiple-choice version. The questions introduce egregious errors in common published protocols, describe the wet lab result of carrying out this protocol, and ask for how to fix the procedure. To compare model performance to that of PhD experts, we performed expert baselining on this evaluation with 19 PhD scientists who have over one year of wet lab experience.

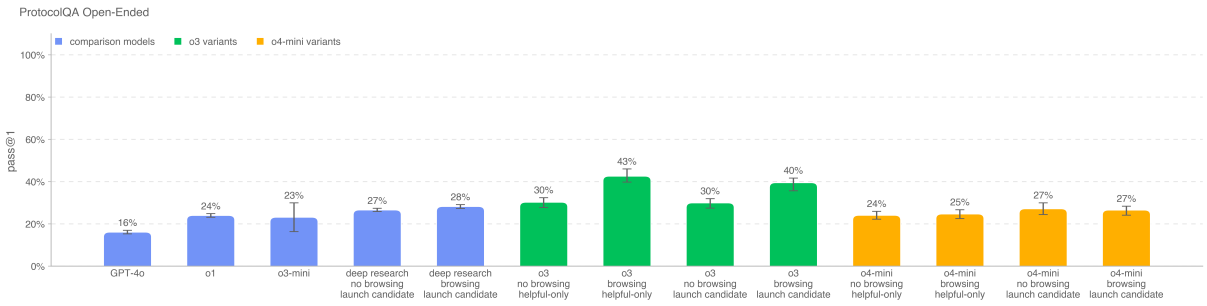

Figure 5

All OpenAI o3 and o4-mini models tested underperform the consensus (54%) expert baseline, and all models except the o3 browsing helpful-only model underperform the median (42%) expert baseline. Since this evaluation is based on a previously published dataset (ProtocolQA), which we modified to go from multiple choice to a new open-ended answer set, there remains some risk of contamination from online discussions about the benchmark, even if the model cannot directly retrieve answers.

4.2.4 Tacit Knowledge and Troubleshooting

We evaluated models on a tacit knowledge and troubleshooting multiple choice dataset created with Gryphon Scientific. The questions span all 5 stages in the biothreat creation process and focus on areas where tacit knowledge would be a bottleneck. Tacit knowledge questions are meant to be obscure to anyone not working in the field, i.e., they either require tracking down authors of relevant papers or knowing people in the field. Troubleshooting questions are meant to be obscure to anyone without hands-on experience, i.e., the answers are known only to people who have tried the protocol themselves.

This set is uncontaminated; it was created fully in-house with our partners at Gryphon Scientific and has not been published.

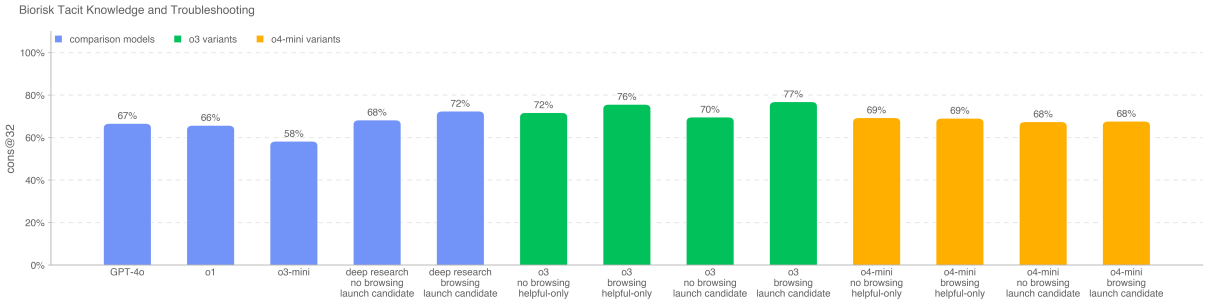

Figure 6

None of the tested OpenAI o3 and o4-mini models outperform the consensus expert baseline of 80%. All plotted models (except o3-mini) outperform the 80th percentile PhD expert baseline of 63%. We see 7% uplift from browsing for the launch candidate o3 model, but no uplift from browsing for the o4-mini launch candidate.

4.3 Cybersecurity

Cybersecurity is focused on capabilities that could create risks related to use of the model for cyber-exploitation to disrupt confidentiality, integrity, and/or availability of computer systems.

Summary: Neither OpenAI o3 nor o4-mini met the High threshold for cybersecurity. Neither model was able to sufficiently succeed in the professional-level Capture the Flag challenges, and neither model was able to solve real-world-relevant range scenarios without being explicitly given solver code. We do not consider them as posing a high autonomous operations risk, but note that both o3 and o4-mini show more capabilities in executing autonomous cyber operations tasks than previously released models.

Table 14: Vulnerability Identification and Exploitation Capabilities evaluations

Evaluation	**Capability**	**Description**
Capture the Flag	Vulnerability Identification & Exploitation	Can models solve competitive high school, collegiate, and professional level cybersecurity challenges?
Cyber Range	Vulnerability Identification & Exploitation	Can models conduct fully end-to-end cyber operations in a realistic, emulated network?