Building an early warning system for LLM-aided biological threat creation View data

We're developing a blueprint for evaluating the risk that a large language model (LLM) could aid someone in creating a biological threat.

In an evaluation involving both biology experts and students, we found that GPT-4 provides at most a mild uplift in biological threat creation accuracy. While this uplift is not large enough to be conclusive, our finding is a starting point for continued research and community deliberation.

Overview
Note: As part of our Preparedness Framework, we are investing in the development of improved evaluation methods for AI-enabled safety risks. We believe that these efforts would benefit from broader input, and that methods-sharing could also be of value to the AI risk research community. To this end, we are presenting some of our early work—today, focused on biological risk. We look forward to community feedback, and to sharing more of our ongoing research.

Background. As OpenAI and other model developers build more capable AI systems, the potential for both beneficial and harmful uses of AI will grow. One potentially harmful use, highlighted by researchers and policymakers, is the ability for AI systems to assist malicious actors in creating biological threats (e.g., see White House 2023(opens in a new window), Lovelace 2022(opens in a new window), Sandbrink 2023(opens in a new window)). In one discussed hypothetical example, a malicious actor might use a highly-capable model to develop a step-by-step protocol, troubleshoot wet-lab procedures, or even autonomously execute steps of the biothreat creation process when given access to tools like cloud labs(opens in a new window) (see Carter et al., 2023(opens in a new window)). However, assessing the viability of such hypothetical examples was limited by insufficient evaluations and data.

Following our recently shared Preparedness Framework(opens in a new window), we are developing methodologies to empirically evaluate these types of risks, to help us understand both where we are today and where we might be in the future. Here, we detail a new evaluation which could help serve as one potential "tripwire" signaling the need for caution and further testing of biological misuse potential. This evaluation aims to measure whether models could meaningfully increase malicious actors' access to dangerous information about biological threat creation, compared to the baseline of existing resources (i.e., the internet).

To evaluate this, we conducted a study with 100 human participants, comprising (a) 50 biology experts with PhDs and professional wet lab experience and (b) 50 student-level participants, with at least one university-level course in biology. Each group of participants was randomly

assigned to either a control group, which only had access to the internet, or a treatment group, which had access to GPT-4 in addition to the internet. Each participant was then asked to complete a set of tasks covering aspects of the end-to-end process for biological threat creation.A To our knowledge, this is the largest to-date human evaluation of AI's impact on biorisk information.

Findings. Our study assessed uplifts in performance for participants with access to GPT-4 across five metrics (accuracy, completeness, innovation, time taken, and self-rated difficulty) and five stages in the biological threat creation process (ideation, acquisition, magnification, formulation, and release). We found mild uplifts in accuracy and completeness for those with access to the language model. Specifically, on a 10-point scale measuring accuracy of responses, we observed a mean score increase of 0.88 for experts and 0.25 for students compared to the internet-only baseline, and similar uplifts for completeness (0.82 for experts and 0.41 for students). However, the obtained effect sizes were not large enough to be statistically significant, and our study highlighted the need for more research around what performance thresholds indicate a meaningful increase in risk. Moreover, we note that information access alone is insufficient to create a biological threat, and that this evaluation does not test for success in the physical construction of the threats.

Below, we share our evaluation procedure and the results it yielded in more detail. We also discuss several methodological insights related to capability elicitation and security considerations needed to run this type of evaluation with frontier models at scale. We also discuss the limitations of statistical significance as an effective method of measuring model risk, and the importance of new research in assessing the meaningfulness of model evaluation results.