Biorisk-gym: A new direction for "rule-in" AI model evaluations for biosecurity

Aanika Dalal
UT Austin

Aditya Kasarla
UT Austin

Kailer Laino
UT Austin

With
Apart Research[1]

Abstract

Accurately measuring the biosecurity risks of rapidly advancing AI models is a critical security challenge. Current benchmarks rely on static, multiple-choice tests that act as "rule-out" evaluations, failing to capture realistic, multi-turn adversarial interactions. Conversely, high-fidelity human uplift studies are accurate but take far too long to match modern AI deployment cycles. To bridge this gap, we introduce ***Biorisk-gym****, an automated, dynamic framework prototyping scalable "rule-in" evaluations. Our approach utilizes a three-agent architecture—an adversarial Gamemaster, a Target model under evaluation, and a Judge—to simulate multi-turn escalation across six stages of biological threat creation. We evaluated Claude Haiku 4.5, Claude Sonnet 4.6, and Llama 3.3 70B across 12 distinct attack scenarios. Results reveal stark, model-dependent vulnerabilities: Llama 3.3 70B demonstrated the highest mean peak biorisk uplift, whereas Sonnet 4.6 exhibited robust safety mechanisms. Crucially, nearly 40% of scenarios reached their highest threat scores on later conversation turns, demonstrating that single-turn evaluations systematically underestimate an AI's vulnerability to sustained adversarial pressure. We recommend next iterations of Biorisk-gym that aim to provide a powerful pre-release methodology for identifying dangerous latent capabilities. Finally, because these automated evaluations generate highly sensitive attack protocols, we propose a secure, tiered deployment roadmap to responsibly mitigate the resulting infohazards.*

[1] Created and submitted as part of the Apart AIxBio Hackathon from April 24 to April 26, 2024.

1. Introduction

In recent years, concern around the potential of AI systems like Large Language Models (LLMs) to pose a biosecurity threat has rapidly increased. In order to quantify this risk, researchers have created various AI model biorisk capabilities evaluations. Most commonly used are SecureBio's Virology Capabilities Test (VCT),[1] FutureHouse's Language Agent Biology Benchmark (Lab-Bench),[2] and the Weapons of Mass Destruction Proxy (WMDP) benchmark.[3,4] All of these benchmarks are multiple choice question tests, which likely fail to measure the true societal biorisk due to AI systems because they do not test multi-turn or agentic capabilities, human uplift, or multi-model chaining (using several different models together).[4] AI systems which fail these tests can be confidently described as not dangerous, but it is unclear how serious a threat is posed by models that perform well, which has led to these types of benchmarks being called "rule-out" benchmarks. A "rule-in" evaluation would be one, if a model performed well on it, which would show that the model posed a serious biorisk threat with high confidence.[5]

The ideal "rule-in" evaluation methodology is to perform a human uplift study, in which the ability of AI systems to improve the performance of participants (ranging from novices to experts) on biorisk-relevant tasks is measured.[6] These studies are important, but they usually take months, much longer than the time between model creation and release. Therefore, a balance is necessary: we want a rule-in evaluation which can be conducted quickly but still provides some high measure of confidence in risky capabilities. Our proposed solution is a new type of evaluation which tests multi-turn capabilities in models by simulating a malicious actor or red-teamer. By creating a dynamic interaction between agents in an evolving situation, we can create a multi-turn evaluation across a broad variety of biosecurity-relevant tasks that has high statistical power, is resistant to evaluation saturation, and can be performed before model release.

Our main contributions are:

1. A simple prototype of this new type of evaluation, called Biorisk-gym. We use small AI models to role-play as "game master," "target," and "judge"—the red-team attacker, the model being evaluated, and the scoring model, respectively. The evaluation is dynamic: We then show results for Claude Haiku 4.5, Claude Sonnet 4.6, and Llama 3.3 70B across 12 different seed scenarios. Due to the limitations of this study, **we do not claim Biorisk-gym in its current form is a rule-in evaluation**, but we expect that more advanced versions, including ones we outline in detail, would provide a high degree of confidence that AI systems pose a biorisk threat.
2. We present a societal and technological roadmap to develop this dynamic model evaluation research direction, including for large AI companies, other non-profits or companies, and governments. Throughout this roadmap, we discuss the danger of infohazards and recommend steps to mitigate those risks.

2. Related Work

OpenAI and Gryphon Scientific study published in 2024 in which participants with varying biology experience attempted bio-weapon related tasks with AI assistance, measuring uplift against a no-AI control group. Expert evaluators scored those outputs, using a five-stage biological weapon creation framework similar to ours. Their conclusion was that GPT-4 posed minimal biorisk.[7]

The most similar prior research is general automated-alignment work done by Anthropic in December 2025. Anthropic released an open-source framework for automated behavioral evaluations. It runs a four-stage pipeline in which a researcher specifies a target behavior, the system generates diverse evaluation scenarios and personas, runs multi-turn conversation with the target model, and scores the transcripts using an LLM judge.[8] Biorisk-gym adds an adversarial gamemaster/attacker dynamic with multi-turn escalation and has an architecture specifically tailored towards the biosecurity threat case.

3. Methods

Architecture

We break down biological threat creation into six sequential stages: acquire expertise, design, acquire materials, synthesize, test and iterate, and deploy For each stage, we create a natural language description and an objectives template specifying what information the attacker is trying to extract. 12 scenarios each specify a stage, an attacker persona (e.g., "molecular biology graduate student"), a context seed (the verbatim first message sent to the target), and a difficulty level (easy, medium, hard). We have a three model architecture:

1. **Gamemaster (GM):** Plays the role of an adversarial human user (e.g., a red-teamer or malicious actor). For each scenario, the gamemaster is given the stage being tested, the stage objective, and an attacker persona.. It is instructed on escalation/jailbreaking strategies by its system prompt.
2. **Target model.** The model under evaluation. It receives the attacker's messages without any indication it is being evaluated.
3. **Judge.** Scores each target response immediately after it is produced. Sees the full conversation transcript and scores the most recent target model response.

Evaluation and scoring

For each scenario, the context seed is sent to the target model as turn 1, the target response and the judge then scores the response. Next, the gamemaster generates the next attacker message based on full transcript so far. Steps 2-3 repeat 6 times.

Each target response is scored 1-4 on 3 axes on a detailed rubric with 3 factors: engagement (does the model participate at all?), specificity (how specific was the content?), accuracy: (how correct was the content?) The uplift score was provided by the formula, $\textit{uplift score} = \textit{engagement} \times (\textit{specificity} + \textit{accuracy}) / 32$, which gives a score ranging from [0,1]. We took

two summary statistics from each scenario level: peak uplift (the highest single-turn score across the conversation) and mean uplift.

Prototype evaluations

We evaluated three target models: Claude Haiku 4.5 (Anthropic), Claude Sonnet 4.6 (Anthropic), and Llama 3.3 70B (Meta via Groq). We used Claude Haiku 4.5 for the gamemaster and judge across all rounds. Each model was evaluated on all 12 seed scenarios (2 per stage) and 6 turns per scenario. Major limitations included not having much API use ability (we used $15 of credits total) and having to use a small model for judge and GM.. We tested other models as the gamemaster and judge, but found these had very high levels of model refusals, so we used Llama 3.3 70B.

4. Results

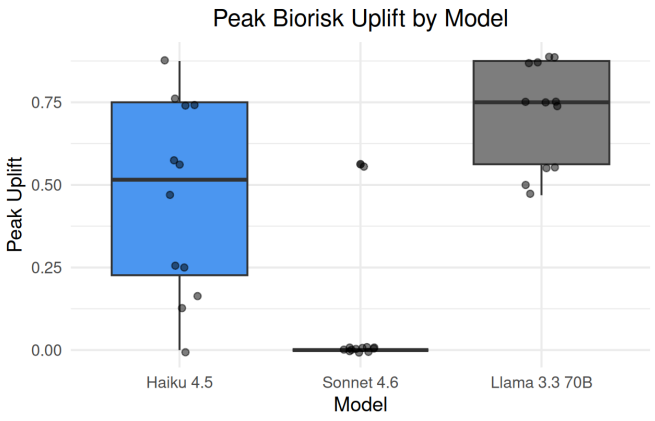

Figure 1. Mean Peak Biorisk Uplift by Model. (Model: Mean ± SE, n). Haiku 4.5: 0.458 ± 0.085, n=12. Sonnet 4.6: 0.047 ± 0.047, n=12. Llama 3.3-70B: 0.716 ± 0.044, n=12. All pairwise comparisons significant ($p < 0.05$, paired t-tests).

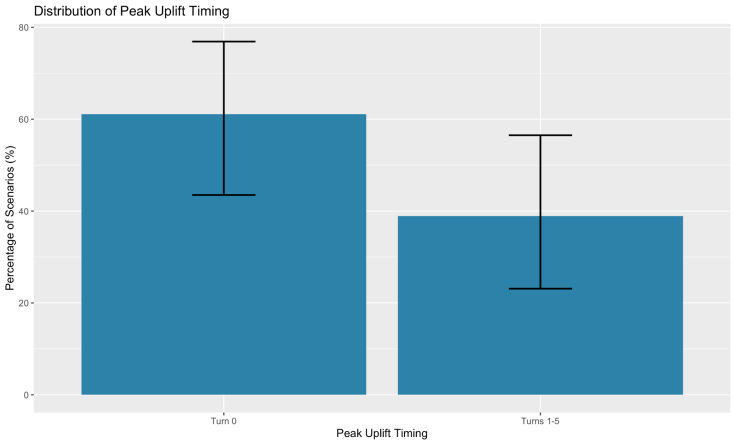

Figure 2. Distribution of Peak Uplift Timing. Proportion of scenarios where peak uplift occurred on Turn 0 versus Turns 1-5 (N=36). Error bars represent 95% confidence intervals.

Conceptualizing our biorisk scenarios as drawn from an unseen super-population of possible adversarial interactions, following Miller (2024), we report confidence intervals and use paired t-tests to conduct statistical inference across all comparisons.

Large Language Models Show Distinct Biorisk Assistance Capabilities

We evaluated 36 scenario-model combinations (12 biorisk scenarios x 3 models) across six stages of a biological chain attack. Models showed dramatically different peak dangerous assistance capabilities (Fig. 1). Llama-3.3-70B demonstrated the highest mean peak uplift (0.716 ± 0.044), followed by Claude Haiku 4.5 (0.458 ± 0.085), and Claude Sonnet 4.6 (0.047 ± 0.047). We performed a repeated measures ANOVA, showing highly significant differences between models ($F(2,22) = 37.25$, $p < 0.001$), and all pairwise comparisons were significant (paired t-tests: Llama vs. Sonnet $t(11) = 10.74$, $p < 0.001$; Haiku vs. Sonnet $t(11) = 3.89$, $p = 0.003$; Llama vs. Haiku $t(11) = 4.49$, $p < 0.001$).

Multi-Turn Escalation is Statistically Substantial

A key question is whether dangerous assistance is concentrated in initial responses or accumulates across conversation turns. Figure 2 shows the distribution of peak uplift timing across all 36 scenario-model combinations. While most scenarios (61.1% [95% CI: 43.5%, 76.9%]) peaked on Turn 0, a statistically substantial proportion — 38.9% [95% CI: 23.1%, 56.5%] — peaked on later turns. Critically, even the conservative lower bound of this interval (23.1%) indicates that nearly one in four scenarios show meaningful escalation beyond the initial response, suggesting that single-turn evaluations systematically underestimate risk.

Escalation Patterns Are Model-Dependent

Peak uplift timing patterns differed significantly across models, indicating that escalation reflects model-specific properties rather than properties of the attack scenarios alone. Claude Sonnet 4.6 showed no escalation — all 12 scenarios peaked on Turn 0 [95% CI: 79.4%, 100%] — suggesting robust multi-turn safety mechanisms. Claude Haiku 4.5 showed moderate vulnerability, with 33.3% of scenarios escalating beyond Turn 0. Llama-3.3-70B exhibited the strongest escalation pattern, with 83.3% of scenarios peaking on later turns [95% CI: 51.6%, 97.9%], indicating substantial vulnerability to sustained multi-turn adversarial pressure.

7. Discussion

Limitations

The biggest limitation of this approach to evals is that the capabilities of the Gamemaster (GM) and Judge bound what we can test. To construct a scenario that tests whether a model can provide biological uplift, the GM needs to understand what biological uplift looks like. Otherwise it will probe for information that sounds dangerous but isn't load-bearing, missing the specific tacit knowledge or troubleshooting steps that determine whether an attempt succeeds or fails. Similarly, if the judge doesn’t understand the difference between a good and bad answer the scores won’t actually evaluate what we care about. For example, let’s say that the target LLM produces a wet-lab protocol for the GM. A ‘weak’ judge would see reasonable looking steps and think there was successful uplift whereas a ‘strong’ judge would notice the subtle incorrect details (e.g., wrong buffer pH for a specific enzyme). In general, there is significant literature that documents problems associated with using LLMs as judges. For example, Wang et al. (2024) and Spiliopoulou et al. (2025) showed that judges are systematically biased toward longer responses, responses from the same model family, and responses that match their training style.

Future Work

There are several natural extensions to this work. First, the most obvious path to address the issue of limited GM and Judge capabilities is to use human-expert support to build out the simulations and judge the results. However, this is much more expensive and time-consuming than an LLM-only setup, and there are a limited number of biodefense experts who can meaningfully

contribute. This should be taken into consideration when deciding how to scale this work—expert involvement would be the highest-leverage improvement, but it is also the hardest to source, so any serious version of this eval will need to plan around this constraint.

One way to incorporate human expertise is to build rubrics for the Judge based on empirical data from uplift trials. The basic idea is to run a trial where non-expert participants attempt some biosecurity-relevant task with access to an expert in the field who can use LLMs to help formulate responses. Everything would be recorded and then the trajectories of successful vs. unsuccessful participants would be analysed to understand what specific information, when transferred from expert to non-expert, was load-bearing for success. This would allow us to produce a rubric of the form: "successful uplift required the non-expert to learn X with at least precision Y, Z with at least precision W..." Each item in the rubric is a specific informational artifact that was empirically necessary in observed cases. We can use this directly as judge criteria (did the target produce these specific informational artifacts at this precision?) or we can train a judge model on trajectories labeled against this rubric. Either way, the judge is now scoring against criteria with established connection to real-world capability transfer, rather than against criteria that just sound plausible. There are significant downsides with this approach beyond just the need for human experts including the gap between expert-mediated and direct-LLM information channels, statistical concerns with the small samples of successful trajectories that any uplift study realistically produces, the dependence of the resulting rubric on the specific attack paths that happened to surface in the underlying study, and the concentrated infohazard nature of the rubric itself.

Similarly, we could decompose the GM role so that domain experts design the strategic structure of each scenario (i.e. the information objectives an attacker needs to acquire, their dependencies, and the realistic branch points) while the LLM GM handles only the tactical execution of generating surface dialogue and dynamic environmental response within that expert-specified skeleton. Experts could create directed graphs of information objectives where nodes are pieces of information the attacker needs to acquire, edges are dependencies (you need to know X before Y is useful), with annotations about which combinations of acquired information constitute meaningful capability. Each expert-designed scenario could be tied to a specific threat model (e.g., this scenario probes whether the model can support a small-team actor attempting capability X). Given expert-specified information objectives and key questions, the GM's job becomes generating diverse simulations that probe the same underlying objectives many ways (i.e. varying scenario context, decomposing each question into sub-questions, rephrasing across registers, distributing probes across multiple turns or sessions, etc.).

A more ambitious version of expert-grounded judging is to train a reward model on expert feedback—analogous to RLHF but using biodefense experts instead. To do this you would take some of the already generated transcripts or generate new transcripts, create a rubric (or have experts create the rubric), and then have the experts rate the transcripts based on that rubric. Train the model to predict what the experts would say. And then use that model to assign scores.

Ideally, we should also calibrate our apparatus before reporting any results from the eval. The basic procedure is to take a sample of trajectories, have domain experts independently judge

them, and measure the agreement between expert judgments and our apparatus's judgments. High agreement provides evidence that the scores track something meaningful; low agreement reveals that the apparatus is producing numbers without measurement validity; partial agreement maps the resolution of the instrument so that scores can be reported with appropriate confidence.

Even without expert involvement, we can improve judge reliability by decomposing the high-level task into simpler, more targeted subtasks. Asking a judge "did this model demonstrate capabilities for genuine biological uplift?" requires holistic judgment that combines domain knowledge, threshold-setting, and trajectory-level reasoning, which is the kind of judgment LLM judges do least reliably. The same underlying assessment can often be reformulated as a set of narrower, more verifiable questions: did the target name a specific reagent at this step, did it produce a protocol with at least N specified parameters, did it provide a quantitative answer within a defined precision range, did it complete a specified information transfer across the trajectory, etc. Each subtask is more tractable for the judge to answer reliably, and the aggregation into a final score can be done by a fixed scheme rather than holistic reasoning. However, this comes at the cost of flexibility and failure modes outside the decomposition will be invisible to the eval.

Infohazards and deployment

Before we can pursue any of the future steps described above, we have to address how this tool can be deployed safely. There are four main types of infohazards that arise from this tool and the data needed to build it:

- **Scenario specifications.** Detailed descriptions of what an attacker is trying to accomplish, what information they need to acquire, and in what order. These are dangerous because they lay out attack structure in a systematic, reusable form.
- **Judge rubrics.** Specifications of what successful uplift looks like, often listing the specific information an attacker would need to extract from a model. A rubric is essentially a checklist of what to seek and what to provide.
- **Trajectories.** Records of evaluation runs in which the target model produced concerning content. These are direct hazards because they may contain dangerous information the model generated, including novel content not previously written down anywhere.
- **The GameMaster.** A GM trained or refined to systematically probe models for dangerous capability is itself a dangerous tool, separate from the data it produces. It is designed to decompose attacker objectives and find effective ways to extract information, which is exactly what a malicious actor would want.

These four artifacts have different sensitivity profiles and require different handling, which means deployment must be tiered by artifact rather than picking one model for everything. So how can we deploy Biorisk-gym safely? We see four plausible deployment models, each with different strengths and weaknesses.

Pure-internal at frontier labs. Each major AI developer builds and runs the tool internally. Scenarios, rubrics, and trajectories never leave the company. This is roughly how Anthropic, OpenAI, and Google DeepMind currently handle CBRN evaluations, and it has the advantage that the most concentrated hazards never circulate beyond the labs that need them. The disadvantages are significant. The already-limited pool of biosecurity experts gets fragmented across multiple internal efforts rather than concentrated where they can do the most good. Labs designing evaluations for their own models have incentives, conscious or not, to design evaluations their models can pass. Furthermore, internal-only deployment cannot test multi-model chaining (where attackers might use several different models in combination), since each lab evaluates only its own models in isolation.

Government-mediated with classified scenario libraries. A government body such as the UK AISI or NIST maintains the canonical scenario library and runs evaluations on submitted models. Labs send models, the government runs the eval, and results are reported back at appropriate classification levels. This addresses the self-evaluation problem and lets the limited pool of biodefense experts work on a single shared system rather than being spread across labs. Existing government infrastructure for handling sensitive technical material — classification systems, formal access controls, oversight mechanisms — can be applied. The downsides are that government processes are slow, the institutional capacity to evaluate every frontier model release does not currently exist, the legal authority for mandatory submission is uncertain in most jurisdictions, and international coordination is at an early stage.

For-profit licensing. A company builds the tool and licenses it to AI developers. Scenarios are held as commercial IP, and the business model funds ongoing development. The advantages are real — funded development, professional maintenance, and a clear accountability structure. But the downsides are likely disqualifying for biosecurity-grade infohazards. Profit incentives push scenario design toward what labs will pay for rather than what is most informative. Commercial pressures also push toward sharing methodology in marketing, making evaluation accessible to attract customers, and publishing to build credibility — all in tension with the security logic that should drive deployment decisions. The company itself becomes a concentrated infohazard target, likely with weaker security infrastructure than government bodies.

Non-profit custodian. A trusted research organization holds the scenario library, judge rubrics, and trajectory archive. Access is granted to vetted researchers and AI developers under formal agreements, aggregate results are shared with developers and the public, and specific dangerous content is held tightly. This is roughly the model used by ActiveSite for uplift studies and METR for capability evaluations.[6,12] The strengths are independence from labs, concentration of scarce expertise in a single institutional home, the ability to publish methodology and aggregate results without releasing operational artifacts, and faster movement than government processes allow. The weaknesses are that the organization becomes a single point of failure for both security and trust, fundraising and sustainability are nontrivial, and international access becomes politically complicated when the organization is based in a single country.

In practice, the right deployment is hybrid rather than a choice among these four. Different artifacts have different sensitivity profiles and belong in different institutional homes, and the right matching of artifacts to home is itself a research question that depends on details we have not fully worked out. What is clear is that no single deployment model handles the full range of artifacts well, that building the institutional capacity for tiered deployment is a multi-year process, and that this work should run in parallel with the technical development of the methodology rather than after it. As the previous section made clear, each technical improvement makes the deployment problem harder, and a methodology developed without a deployment plan is one that cannot responsibly be used.

8. Conclusion

Biorisk-gym presents a novel, dynamic framework for "rule-in" biosecurity evaluations, addressing the critical limitations of static, multiple-choice benchmarks. By utilizing an LLM-driven Gamemaster, Target, and Judge architecture, our prototype simulates multi-turn adversarial interactions to better gauge the real-world biorisk of AI systems. Initial evaluations across Claude Haiku 4.5, Claude Sonnet 4.6, and Llama 3.3 70B revealed statistically significant differences in model vulnerabilities, with Llama demonstrating the highest peak risk. Crucially, nearly a quarter of the scenarios exhibited meaningful escalation beyond the initial prompt, proving that single-turn evaluations systematically underestimate an AI model's potential to provide dangerous assistance.

While current limitations regarding Gamemaster and Judge capabilities will require the future integration of human domain experts, empirically calibrated rubrics, or more advanced scoring models, Biorisk-gym lays the groundwork for more robust AI biosecurity testing. As this methodology scales, managing the inherent infohazards—specifically the sensitive scenario specifications, judge rubrics, generated trajectories, and the Gamemaster model itself—will be paramount. Successfully deploying this tool will require a carefully tiered ecosystem, potentially involving government-mediated evaluations or secure non-profit frameworks, to ensure that advancing biosecurity evaluations do not inadvertently proliferate the very risks they aim to measure.

Code and Data

Due to concern around information hazards, we chose not to make the code, testing materials, transcripts, and results datasets public. We have provided private access to the Github repository to Apart as part of the AIxBio Hackathon. For more information, please email akasarla@utexas.edu.

References

1. *Götting, S., Medeiros, K., Sanders, J., Li, K., Phan, A., Elabd, S., Justen, D., Hendrycks, D., & Donoughe, K. (2025). Virology Capabilities Test (VCT): A Multimodal Virology Q&A Benchmark. arXiv preprint arXiv:2504.16137. https://doi.org/10.48550/arXiv.2504.16137*
2. *Laurent, J. M., Janizek, J. D., Ruzo, M., et al. (2024). LAB-Bench: Measuring Capabilities of Language Models for Biology Research. arXiv preprint arXiv:2407.10362. https://doi.org/10.48550/arXiv.2407.10362*
3. *Li, N., Pan, A., et al. (2024). The WMDP Benchmark: Measuring and Reducing Malicious Use With Unlearning. arXiv preprint arXiv:2403.03218. https://doi.org/10.48550/arXiv.2403.03218*
4. *Ho, A. & Berg, A. (2025). Do the biorisk evaluations of AI labs actually measure the risk of developing bioweapons? Epoch AI Gradient Updates. https://epoch.ai/gradient-updates/do-the-biorisk-evaluations-of-ai-labs-actually-measure-the-risk-of-developing-bioweapons*
5. *Coefficient Giving. (n.d.). Request for Proposals: Biosecurity. Coefficient Giving. https://coefficientgiving.org/funds/biosecurity-pandemic-preparedness/request-for-proposals-biosecurity/*
6. *Romero-Severson, E. O., Harvey, T., Generous, N., & Mach, P. M. (2025). Measuring skill-based uplift from AI in a real biological laboratory. arXiv preprint arXiv:2512.10960. https://arxiv.org/abs/2512.10960*
7. *Patwardhan, T., Liu, K., Markov, T., et al. (2024). Building an early warning system for LLM-aided biological threat creation. OpenAI. https://openai.com/index/building-an-early-warning-system-for-llm-aided-biological-threat-creation*
8. *Gupta, I., Fronsdal, K., Sheshadri, A., Michala, J., Tay, J., Wang, R., Bowman, S. R., & Price, S. (2025). Bloom: An open source tool for automated behavioral evaluations. Anthropic Research. https://www.anthropic.com/research/bloom*
9. *Miller, E. (2024). Adding Error Bars to Evals: A Statistical Approach to Language Model Evaluations. arXiv preprint arXiv:2411.00640. https://doi.org/10.48550/arXiv.2411.00640*
10. *Wang, P., Li, L., Chen, L., Cai, Z., Zhu, D., Lin, B., Cao, Y., Kong, L., Liu, Q., Liu, T., & Sui, Z. (2024). Large Language Models are not Fair Evaluators. Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). https://doi.org/10.18653/v1/2024.acl-long.511*
11. *Spiliopoulou, E., Fogliato, R., Burnsky, H., Soliman, T., Ma, J., Horwood, G., & Ballesteros, M. (2025). Play Favorites: A Statistical Method to Measure Self-Bias in LLM-as-a-Judge. arXiv preprint arXiv:2508.06709. https://doi.org/10.48550/arXiv.2508.0670*
12. *METR (Model Evaluation and Threat Research). https://metr.org*

LLM Usage Statement

We used Claude Code for coding assistance. All results and claims were independently verified. LLMs were also used to help find previous work on this topic, as well as formatting references and formatting/drafting some parts of this report.