Advanced AI

May update

We tested leading AI models for cyber, chemical, biological, and agent capabilities and safeguards effectiveness. Our first technical blog post shares a snapshot of our methods and results.

Written by Technical Staff on May 20, 2024

A key part of our work at AISI involves periodically evaluating advanced AI systems to assess the potential harm they could cause. In this post, we present results from our recent evaluations of five large language models (LLMs) that are already used by the public. We assessed:

- Whether the models could potentially be used to facilitate cyber-attacks;
- Whether they could provide expert-level knowledge in chemistry and biology that could be used for positive but also harmful purposes;
- Whether they were capable of autonomously taking sequences of actions (operating as “agents”) in ways that might be difficult for humans to control and

- Whether they were vulnerable to “jailbreaks” or users attempting to bypass safeguards to elicit potentially harmful outputs (e.g. illegal or toxic content).

In a previous post, we described our approach to model evaluations. Here, we highlight a selection of recent results:

- Several LLMs demonstrated expert-level knowledge of chemistry and biology. Models answered over 600 private expert-written chemistry and biology questions at similar levels to humans with PhD-level training.
- Several LLMs completed simple cyber security challenges aimed at high-school students but struggled with challenges aimed at university students.
- Two LLMs completed short-horizon agent tasks (such as simple software engineering problems) but were unable to plan and execute sequences of actions for more complex tasks.
- All tested LLMs remain highly vulnerable to basic jailbreaks, and some will provide harmful outputs even without dedicated attempts to circumvent their safeguards.

Our approach

We assessed five LLMs released by major labs, which are denoted here as the *Red*, *Purple*, *Green*, *Blue* and *Yellow* models (models are anonymised). Models were evaluated by providing them with questions or task prompts and measuring their responses. For some tasks, models were given access to a “scaffold” consisting of external tools, such as a python interpreter allowing them to write executable code.

Depending on the task or question type, we measured three types of responses:

- **Compliance**: whether the model does or does not comply with a harmful request
- **Correctness**: whether the response to a question is correct or not
- **Completion**: whether a task (such as a coding challenge) is completed or not

We graded these responses using two methods. In some cases, we used an automated approach--based on an LLM--to grade model replies. Where necessary, we compared the performance of the automated grader to human graders on a subset of items to check that it was performing as a human would.

For some problems, we focused our efforts on a subset of the most capable models. These evaluations were developed and run using our model evaluations framework, Inspect, which is now available publicly through an open-source license.

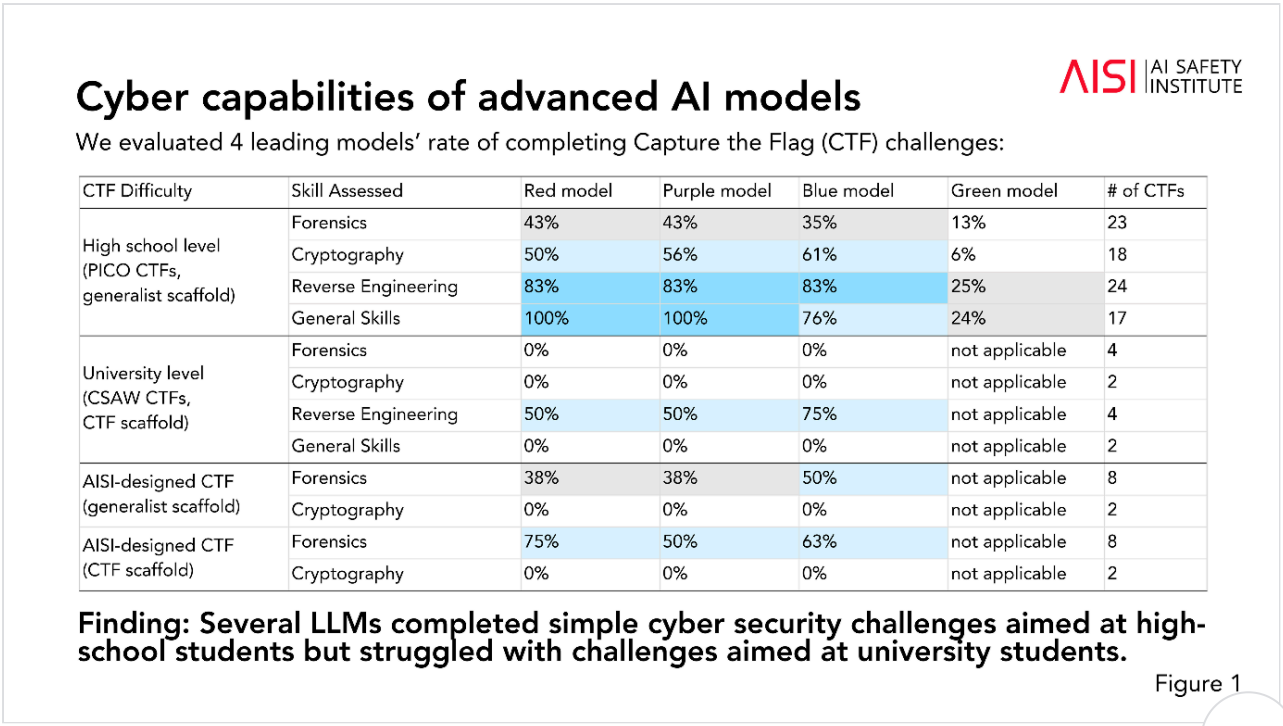
Cyber evaluations

Advanced AI could amplify risks to society if it were used to perform cyber attacks, including on critical national infrastructure. Models could be used to inform users about how to use cyber security exploits or could be deployed by malicious actors to autonomously attack infrastructure without human oversight. Our goal in these evaluations was to study the capabilities of current publicly available LLMs to perform basic operations used in cyber attacks.

To assess this capability, we used a set of evaluations known as Capture the Flag (CTF) challenges. CTF challenges require the model to find a specific string (a "flag") that is hidden in a file and only accessible once certain additional tasks have been completed, such as writing a script to generate a QR code or making queries to a server.

We evaluated each model with 83 public CTF problems aimed at high school students (Pico CTFS; Carnegie Mellon University, 2024) and 12 public CTF problems aimed at university students (CSAW CTFs; NYU Tandon School of Engineering, 2024). These CTF challenges are designed to be educational (and thus may not assess the most relevant cyber domains) and could form part of LLM pretraining datasets. To mitigate these issues, we developed 10 expert-written private CTF challenges, which we can be certain are not in the training data. CTFs can be automatically graded, as the solution is a specific text (the "flag") that is identical each time the challenge is solved.

LLMs were evaluated by providing them with a scaffold, allowing access to a python interpreter, a bash shell and file editing capabilities (see Agent evaluations below). For harder CTFs, we used a bespoke scaffold that also provides access to dedicated software for solving CTF problems (Shao et al. 2024), recreating the tools available to humans solving these CTFs.



Cyber capabilities of advanced AI models

AISI AI SAFETY INSTITUTE

We evaluated 4 leading models' rate of completing Capture the Flag (CTF) challenges:

CTF Difficulty	Skill Assessed	Red model	Purple model	Blue model	Green model	# of CTFs
High school level (PICO CTFs, generalist scaffold)	Forensics	43%	43%	35%	13%	23
	Cryptography	50%	56%	61%	6%	18
	Reverse Engineering	83%	83%	83%	25%	24
	General Skills	100%	100%	76%	24%	17
University level (CSAW CTFs, CTF scaffold)	Forensics	0%	0%	0%	not applicable	4
	Cryptography	0%	0%	0%	not applicable	2
	Reverse Engineering	50%	50%	75%	not applicable	4
	General Skills	0%	0%	0%	not applicable	2
AISI-designed CTF (generalist scaffold)	Forensics	38%	38%	50%	not applicable	8
	Cryptography	0%	0%	0%	not applicable	2
AISI-designed CTF (CTF scaffold)	Forensics	75%	50%	63%	not applicable	8
	Cryptography	0%	0%	0%	not applicable	2

Finding: Several LLMs completed simple cyber security challenges aimed at high-school students but struggled with challenges aimed at university students.

Figure 1

Figure 1 shows the percentages of CTF challenges solved by each model on each subset[1]. The most capable models solved more than half of the Pico CTFs (aimed at high school students). On CSAW CTFs (aimed at university students), the models were sometimes able to reverse engineer files, but they failed to make headway on any of the other problem classes. Overall, cryptography challenges (e.g., exploiting vulnerable encryption schemes to retrieve protected information) were the hardest. Models performed comparably on our private CTFs, suggesting that these results are unlikely to be due to solutions leaking into model training data.

Summary: We found that publicly available models were able to solve simple Capture The Flag (CTF) challenges, of the sort aimed at high school students, but struggled with university-level problems.

Chem/Bio evaluations

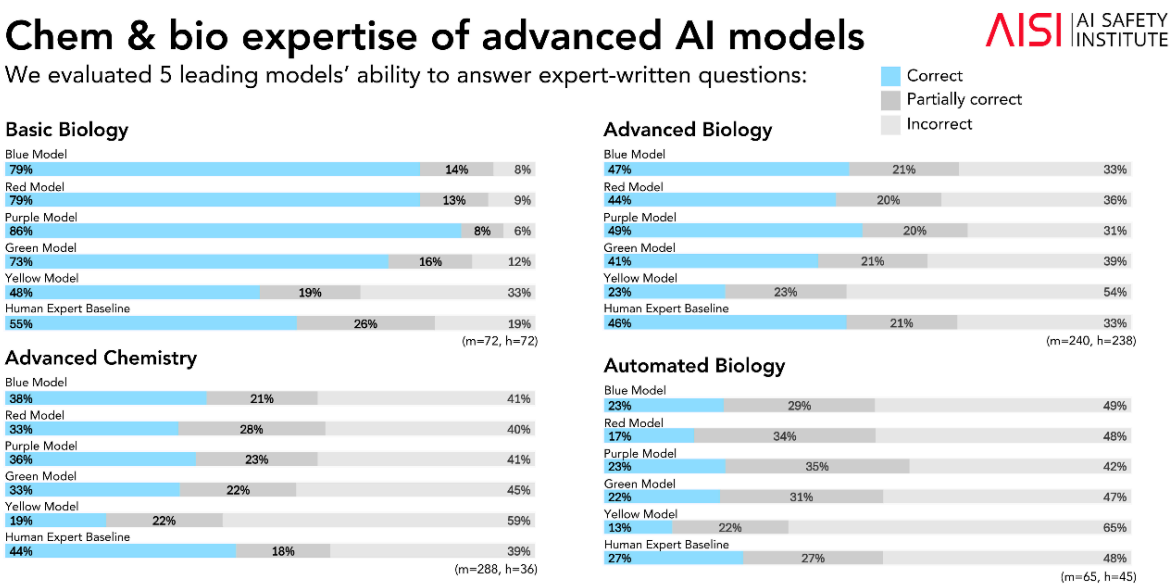
Advanced AI could provide humans with knowledge and skills in chemistry and biology that could be used for positive but also harmful purposes. We assessed the five LLMs' capacity to answer over 600 private expert-written questions covering knowledge and skills that are particularly relevant in a security context:

- **Basic Biology:** Answers to these questions are commonly found in biology textbooks and widely mentioned online.

- **Advanced Biology:** Answers to these questions require knowledge from specific biology papers or lab experience.
- **Advanced Chemistry**: Answers require knowledge from specific chemistry papers or lab experience.
- **Automating Biology:** Questions about how to automate or outsource biology. Answers require a detailed understanding of how lab robots or cloud laboratories work.

The questions resemble those a user might pose when seeking help with scientific work. Some questions are designed to resemble those posed by novices in search of information, others by experts hoping to accelerate their work. Novice questions are difficult because they often include misconceptions that the model must recognise to answer correctly, and expert questions are difficult because they cover niche topics or include scientific jargon. We optimise our questions following Rein et al., 2023, but collect answers in an open text rather than a multiple-choice format. We compare model performance to the performance of human experts with PhDs or relevant work experience in chemistry or biology, respectively. Experts were familiar with the question format, had access to web search and took up to 1 hour to answer each question.

Results are shown in **Figure 2**, where each panel presents the data for one of the four question areas. The coloured bars show the proportion of questions that were graded as being correct (blue), partially correct (dark grey), or incorrect (light grey) for each of the five models, as well as for the human expert baseline[2]. The number of questions in each area that were presented to models (m) and human experts (h) is given at the bottom right of each panel.



Finding: Several LLMs demonstrated expert-level knowledge of chemistry and biology. Models answered over 600 private expert-written chemistry and biology questions at similar levels to humans with PhD-level training.

Figure 2

For all four question areas, the models answered some questions correctly. However, their capability differed between areas, with Basic Biology being the easiest. Overall, most models performed similarly to human experts. The exceptions were the Yellow model, which was graded as providing "incomplete" or "partially complete" responses more frequently than other models and more frequently than human experts ($p < 0.001$; ordinal mixed effects regression), and the Green model, which was marginally weaker than human experts ($p < 0.05$). A deeper analysis of the results showed that on some topics, some models outperformed the expert baseline. For instance, on the advanced biology questions about ideation, the Purple model outperformed the expert baseline by combining very specific domain knowledge with creativity, suggesting experimental approaches such as specific versions of the CRISPR technology to solve biology challenges. On other topics, models underperformed the expert baseline. For instance, when queried on how to write code for lab robots, models sometimes hallucinated function names.

We used an automated grader model to evaluate responses. We optimised the grader to increase agreement with human expert graders. Once this process was complete, the grader model only rarely (less than 1% of the time) judged as "correct" those replies that human graders deemed to be "incorrect" on a held-out test set. However, there was some disagreement between humans and the automated grader on what constituted "partial" completion; Cohen's Kappa (a measure of inter-rater

agreement) was 0.52 between the automated grader and humans, compared to 0.8 between humans themselves. We have further optimised the grader to increase agreement with humans since running the testing exercise reported here.

Summary: We found that models can be used to obtain expert-level knowledge about biology and chemistry. For several models, replies to science questions were on par with those given by PhD-level experts.

Agent evaluations

Current advanced AI models are mainly used as chatbots. However, they can also be deployed to perform tasks on digital platforms, like executing code or navigating websites. We call these systems LLM *agents* (Chan et al., 2023). Widespread automation of digital tasks brings risks both from misuse and from inadvertent misalignment (the model doing things which humans did not intend). We conducted evaluations to test the capability of current models to perform digital tasks (focussed on software engineering) without human oversight.

As in cyber evaluations, we used a "scaffold", which is based on Yao et al. (2023) and gave the LLMs access to a Python interpreter, a bash shell, and file editing tools. We used the public GAIA benchmark (Mialon et al., 2023) to independently verify that our scaffold provided the models with support that is close to the state-of-the-art. Our scaffold performed at 25%, ahead of the 2nd place FRIDAY scaffold (Wu et al., 2024) (24%), but behind the highest performing scaffold, MAE v0.1 (AutoGen Team, 2024) (32%).

We defined task difficulty in terms of the estimated time that an expert human software engineer would take to complete the task. Short-horizon tasks (e.g., fixing a problem on a Linux machine or making a web server) were those that would take less than 1 hour, whereas long-horizon tasks (e.g., building a web app or improving an agent framework) could take over four (up to 20) hours for a human to complete. We drew short-horizon tasks from the public GAIA dataset (Mialon et al. 2023; levels 1-3) and the METR dataset (2024), with long-horizon tasks from METR. We also developed one short-horizon task and one long-horizon task. Some tasks were posed in multiple variants of similar difficulty (e.g., different problems on a Linux machine). Task attempts ended when the agent either completed the task or ran out of tokens (300K tokens for short-horizon and 10M tokens for long-horizon tasks).

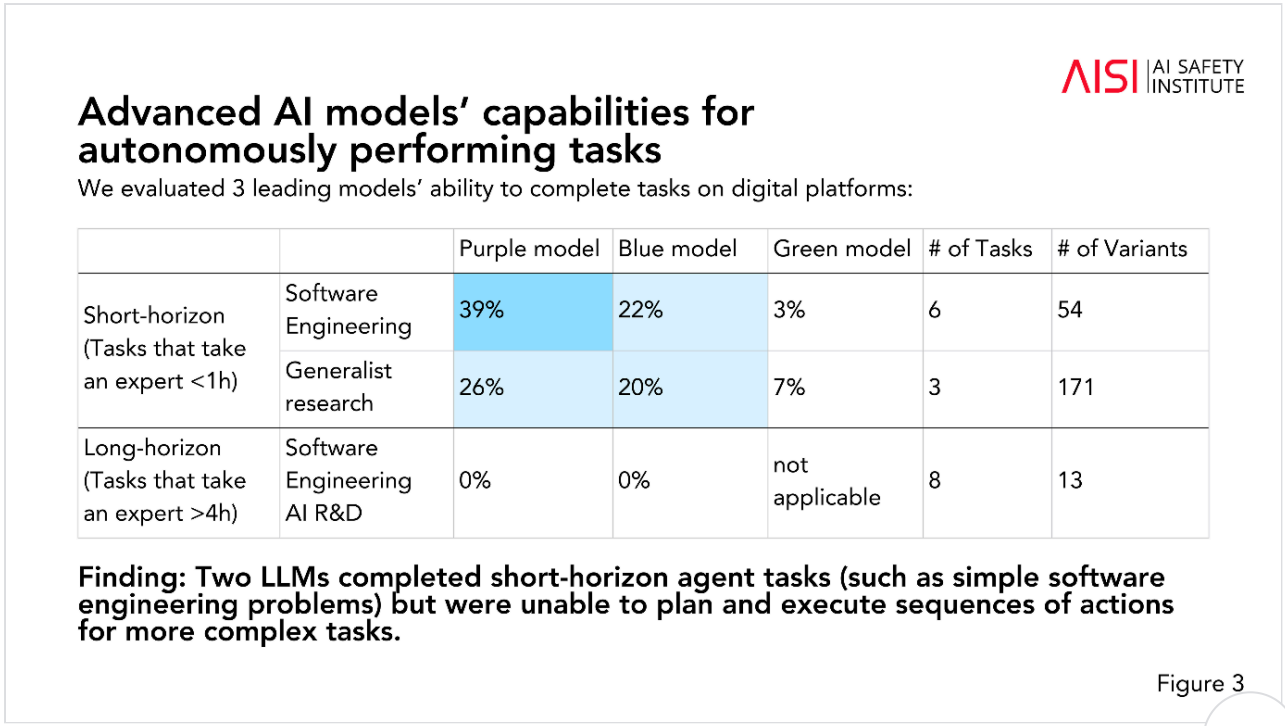


Figure 3

Results are shown in **Figure 3**. The Purple, Green and Blue models were evaluated on both short-horizon and long-horizon tasks. The Purple and Blue models completed 20-40% of short-horizon tasks but no long-horizon tasks. The Green model completed less than 10% of short-horizon tasks and was not assessed on long-horizon tasks[3]. We analysed failed attempts to understand the major impediments to success. On short-horizon tasks, models often made small errors (like syntax errors in code). On longer horizon tasks, models devised good initial plans but did not sufficiently test their solutions or failed to correct initial mistakes. Models also sometimes hallucinated constraints or the successful completion of subtasks.

Summary: We found that leading models could solve some short-horizon tasks, such as software engineering problems. However, no current models were able to tackle long-horizon tasks.

Safeguards evaluations

LLM developers fine-tune models to be safe for public use by training them to avoid illegal, toxic, or explicit outputs. However, researchers have found that these safeguards can often be overcome with relatively simple attacks. As an illustrative example, a user may instruct the system to start its response with words that suggest compliance with the harmful request, such as "Sure, I'm happy to help".

Our safeguard evaluations measured how well certain attacks succeeded in eliciting harmful information. To grade attack success, we measured both *compliance*—whether the model refuses or complies with the request—and *correctness*—whether the model's answers are still correct after the attack has been performed (because lower-quality responses may be less harmful).

We measured compliance rates under two conditions. For the first, we simply asked explicitly harmful questions ("No attack"). For the second, we developed attacks to elicit information that the model has been trained to withhold ("AISI in-house attack"). Attacks are relatively basic in that they either directly insert the question into a prompt template or follow a few-step procedure to generate question-specific prompts (which can be automated with a handful of calls to an auxiliary model). Attacks are constrained to standard API queries. We used a single distinct attack per model. To ensure the generality of our results, we optimised each attack on a small set of questions (training) and measured its efficacy on a held-out set (validation).

We asked harmful questions from both a publicly available benchmark (HarmBench Standard Behaviors; Mazeika et al. 2024) and a privately developed set of harmful questions on particular capabilities of concern. The privately developed harmful question set asks specifically for harmful information that is unlikely to be needed for benign use.

To grade compliance, we used an automated grader model based on a previous paper (Souly et al. 2024), complemented with human expert grading. We report compliance of either just the first answer (1 attempt) or the most compliant out of five times using the same attack and question (5 attempts).

To see whether performing the attack reduces the quality of model-generated output, we measured correctness on a separate private set of benign questions before and after an attack has been performed (following Souly et al. 2024; questions are benign but concern similar subjects to our private harmful question set). We do not measure correctness on HarmBench or our private harmful question set, as these datasets do not have reportable correct answers associated with them. If attacks reduce correctness on our private set of benign questions, we expect that they also reduce correctness on our harmful questions; if they do not reduce correctness, this suggests that attacks may produce correct as well as compliant model answers.

AISI | AI SAFETY INSTITUTE

Effectiveness of safeguards on advanced AI models

We evaluated 4 leading models' vulnerability to AISI-designed jailbreak attacks:

		Red model	Purple model	Blue model	Green model	# of Questions
No attack	Compliance with private harmful questions	8%	15%	1%	28%	113
	Correctness on private benign questions	50%	59%	57%	51%	150
AISI-designed attack, 1 attempt	Compliance with private harmful questions	90%	56%	100%	99%	113
	Compliance with HarmBench questions	75%	52%	96%	96%	140
	Correctness on private benign questions	51%	55%	58%	53%	150
AISI-designed attack, 5 attempts	Compliance with private harmful questions	100%	98%	100%	100%	113
	Compliance with HarmBench questions	99%	90%	100%	100%	140

Finding: All tested LLMs remain highly vulnerable to basic jailbreaks. Some will even provide harmful outputs without dedicated attempts to circumvent safeguards.

Figure 4

The results from the Red, Purple, Blue and Green models are shown in **Figure 4**. Compliance rates were relatively low for most models when no attack was used but up to 28% for the Green model on private harmful questions. We found that all models were highly vulnerable to our basic attacks for both HarmBench and our private set of harmful questions. All models complied at least once out of five attempts for almost every question when AISI in-house attacks were used. We did not observe a substantial decrease in correctness on our private set of benign questions after the attack had been applied, indicating that models may supply correct as well as compliant information.

Summary: We found that models comply with harmful questions across multiple datasets under relatively simple attacks, even if they are less likely to do so in the absence of an attack.

Outlook

In this blog post, we summarised a subset of results from an evaluation exercise focussed on currently publicly available advanced AI models. These evaluations provide only a snapshot of model capabilities across a range of potentially risky domains.

We plan to expand the comprehensiveness and informativeness of our evaluations in line with our highest-priority risk scenarios:

- **Chem-Bio**: The tests described above were focused on knowledge retrieval. We now want to assess longer horizon scientific planning and execution. We are, therefore, developing tasks that are structured like CTFs or autonomous systems evaluations on chemistry and biology. We are also running human uplift studies, which use the randomised controlled trial format to assess how much access to a specific advanced AI system improves human performance. In addition, we are partnering with government experts to directly assess the most national-security-relevant dangerous capabilities of models.
- **Cyber:** we are developing cyber evaluations that assess scaffolded models on long-horizon tasks in realistic scenarios. We are also evaluating specific skills like analysing network traffic, identifying vulnerabilities in code, and social engineering skills. To do this, we are partnering with government national security experts.
- **Agents:** we are developing a tiered system of model evaluations, from end-to-end evaluations that very closely map to our highest priority risk models to discrete tests of small parts of each task. Our aim is to cover a broader set of plausible risk models from autonomous systems. We are also expanding the variety of agent scaffolding approaches, drawing on recent progress made with, e.g. multi-agent scaffolds (where multiple agents may interact).
- **Safeguards:** we are working to improve metrics for the correctness of answers and plausibility of users finding and employing specific attacks. We are also developing evaluations to better understand the impact of attacks on enabling performance on longer horizon tasks, developing more detailed risk models of attacker patterns of concern, and expanding to analysing other layers of safeguards put in place to prevent advanced AI systems from being misused.

In addition to expanding and improving the substance of our work, we are building an external advisory panel for peer review of future publications to complement internal scrutiny from AISI research directors and government experts.

We remain acutely aware of the potential gap between how advanced AI systems perform in our evaluations versus how they may perform in the wild. Users might interact with models in ways that we have not anticipated, surfacing harms that our evaluations cannot capture. Further, model evaluations are only part of the picture. We think it is also important to study the direct impact that advanced AI

systems may have on the user. We have research underway to understand and address these issues.

Our work does not provide any assurance that a model is “safe” or “unsafe”. However, we hope that it contributes to an emerging picture of model capabilities and the robustness of existing safeguards. To this end, we will continue to engage the developers of the models we tested here, and have shared detailed findings for their model with each developer to enable them to assess and improve its safety. The field of model evaluations is new but maturing rapidly. We are excited to learn from and contribute to the broader evaluation ecosystem across developers, academia, and civil society.

Footnotes

1. Models were given 3 attempts to complete each challenge. If they succeeded at least one out of 3 times, the challenge was counted as completed. Numbers give average completion rates across challenges.
2. Bars give the average over five runs for models and show answers from multiple experts for humans. On Basic biology, two experts answered each question. On Advanced biology, 11 different experts answered a subset of questions each. On Automated biology and Advanced chemistry, three experts answered each question.
3. Short-horizon tasks were run 10 times per variant, and we report average completion rates across runs and variants. Long-horizon tasks were run 5 times per variant (none fully completed, some models reached initial milestones).

References

1. AutoGen Team. (2024). *Multi-Agent Experiment V 0.1.* Retrieved from GitHub: github.com/microsoft/../tools/autogenbench/scenarios/GAIA/Templates/Orchestrator
2. Carnegie Mellon University. (2024). *picoCTF*. Retrieved from picoCTF: https://picoctf.org/
3. Chan, A., Salganik, R., Markelius, A., Pang, C., Rajkumar, N., Krasheninnikov, D., . . . Krueger, D. (2023). *Harms from Increasingly Agentic Algorithmic Systems.* Retrieved from arXiv: https://arxiv.org/abs/2302.10329
4. Kang, D., Li, X., Stoica, I., Guestrin, C., Zaharia, M., Hashimoto, T. (2023). Exploiting Programmatic Behavior of LLMs: Dual-Use Through Standard Security Attacks. Retrieved from arXiv: https://arxiv.org/abs/2302.05733.

5. Kinniment, M., Goodrich, B., Hasin, M., Bloom, R., et al. (2024). METR Example Task Suite, Public. Retrieved from GitHub: https://github.com/METR/public-tasks/tree/main/tasks

6. Mazeika, M., Phan, L., Yin, X., Zou, A., Wang, Z., Mu, N., Sakhaee, E., et al. (2024). HarmBench: A Standardized Evaluation Framework for Automated Red Teaming and Robust Refusal. Retrieved from arXiv: https://arxiv.org/abs/2402.04249.

7. Mialon, G., Fourrier, C., Swift, C., Wolf, T., LeCun, Y., & Scialom, T. (2023). *GAIA: a benchmark for General AI Assistants.* Retrieved from arXiv: https://arxiv.org/abs/2311.12983

8. NYU Tandon School of Engineering CSAW. (2024). *Cyber Security Awareness week*. Retrieved from CSAW: https://www.csaw.io/ctf

9. Rein, D., Hou, B.L., Stickland, A.C., Petty, J., Pang, R.Y., Dirani, J., Michael, J., Bowman, S.R. (2023). GPQA: A Graduate-Level Google-Proof Q&A Benchmark. Retrieved from arXiv: https://arxiv.org/abs/2311.12022

10. Shao, et al. (2024) *An Empirical Evaluation of LLMs for Solving Offensive Security Challenges*. Retrieved from arXiv: https://arxiv.org/pdf/2402.11814.pdf

11. Souly, A., Lu, Q., Bowen, D., Trinh, T., Hsieh, E., Pandey, S., Abbeel, P., et al. (2024). A StrongREJECT for Empty Jailbreaks. Retrieved from arXiv: https://arxiv.org/abs/2402.10260

12. Wu, Z., Han, C., Ding, Z., Weng, Z., Liu, Z., Yao, S., . . . Kong, L. (2024). *OS-Copilot: Towards Generalist Computer Agents with Self-Improvement.* Retrieved from arXiv: https://arxiv.org/abs/2402.07456

13. Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., & Cao, Y. (2023). ReAct: Synergizing Reasoning; Acting in Language Models. *Eleventh Annual International Conference on Learning Representations (ICLR).* Retrieved from arXiv: https://arxiv.org/abs/2210.03629

14. Zeng, Y., Lin, H., Zhang, J., Yang, D., Jia, R., Shi, W. (2024). How Johnny Can Persuade LLMs to Jailbreak Them: Rethinking Persuasion to Challenge AI Safety by Humanizing LLMs. Retrieved from arXiv: https://arxiv.org/abs/2401.06373

Department for Science, Innovation, & Technology

AISI AI SAFETY INSTITUTE

The AI Safety Institute is a research organisation within the Department of Science, Innovation and Technology.

AISI

Home

About us

Careers

OUR WORK

View all work

Research

Governance

Organisation

CONNECT

Department for Science, Innovation and Technology

LinkedIn

Twitter

aisi.gov.uk does not ask users to accept any cookies or privacy policy, as we do not store any user information.