NIST CENTER FOR AI STANDARDS AND INNOVATION (CAISI)

Assessment of Z.ai's GLM-5.2

Center for AI Standards and Innovation

National Institute of Standards and Technology

July 8, 2026

1. Executive Summary

GLM-5.2 was released as an open-weight model by the PRC-based company Z.ai (formerly known as Zhipu AI) on June 16, 2026. CAISI evaluated GLM-5.2 and found that:

- **GLM-5.2 was probably the most capable open-weight AI model when it was released (Figure 1.1).** According to CAISI's set of evaluations:
  - GLM-5.2's overall capabilities are similar to that of GPT-5.2, released in December 2025.
  - GLM-5.2's cyber capabilities are similar to that of Opus 4.6, released in February 2026.

- **GLM-5.2's performance on safeguards and security is mixed.** According to CAISI's set of evaluations:
  - GLM-5.2's safeguards allow assistance with agentic cyber exploit development.
  - GLM-5.2's safeguards block fewer sensitive biological questions than reference U.S. models.
  - However, GLM-5.2 appears potentially more robust against agent hijacking and jailbreaking attacks than other evaluated PRC open-weight models.
  - These evaluations measure robustness against prompt-based jailbreaks. Regardless of their robustness, safeguards for open-weight models can be circumvented when self-hosted.

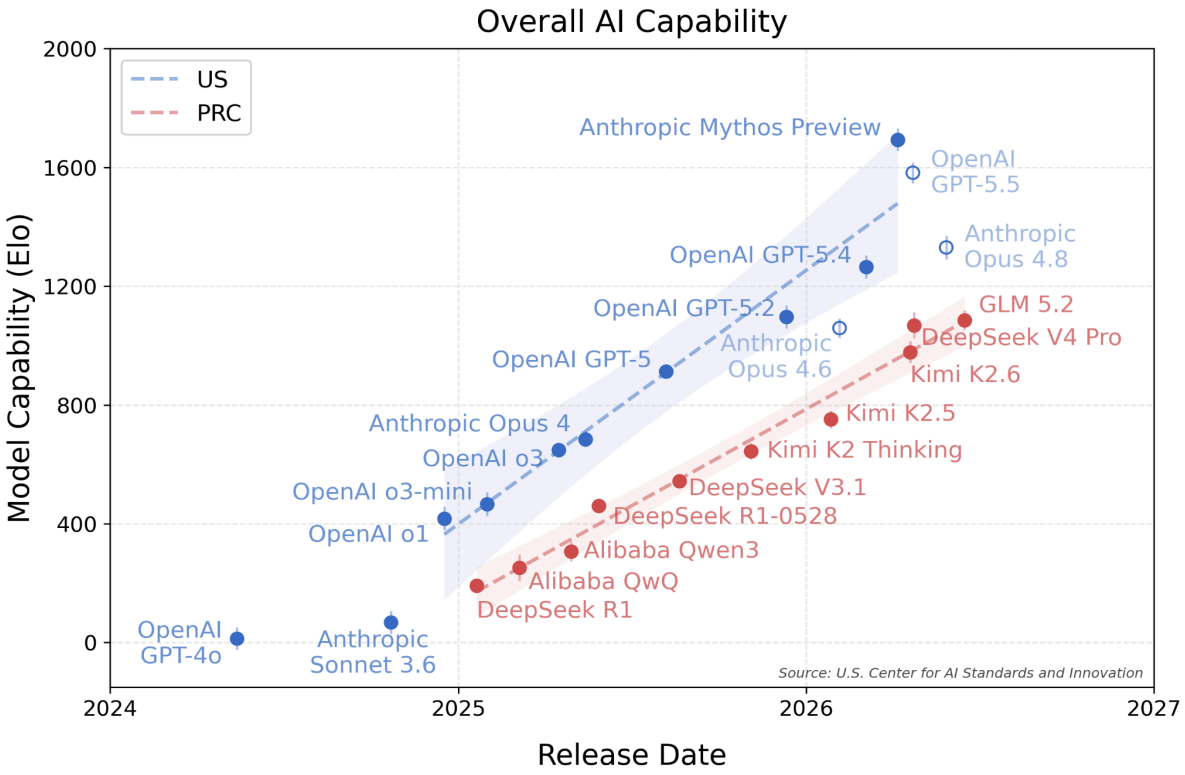

Figure 1.1: Comparison of the overall capabilities over time of the most capable U.S. and PRC models as of GLM-5.2's release, based on CAISI's evaluation results. Open circles denote non-frontier models. See Section 3 on "Overall Capabilities" for more details.

Table of Contents

2. Report Scope and Context

2.1 Scope

- This report aims to:
  - Assess the capabilities, security, and safeguards of GLM-5.2.
  - Compare GLM-5.2 against historical frontier AI models from both the U.S. and PRC.
- Sources for this report include:
  - CAISI's independent evaluations of GLM-5.2.
  - The developer's self-reported evaluation results of GLM-5.2.

2.2 Context

- Z.ai was founded in 2019 (as Zhipu AI) in Beijing by researchers from Tsinghua University.
- Z.ai started releasing the GLM series of large language models in March 2021.
- Z.ai completed its IPO in Hong Kong on January 8, 2026, becoming the first major PRC AI developer to go public. It recently announced plans to list in Shanghai as well to raise more capital. At IPO, Z.ai's market capitalization was $7B, and has now increased by about 10X to $65B (as of July 17, 2026).
- In 2025, Z.ai's revenue was $105M, and its annual revenue has doubled every year since 2022.

3. Overall Capabilities

3.1 Summary

CAISI Overall Assessment	**GLM-5.2 was probably the most capable open-weight AI model when it was released, but it lags behind the U.S. frontier.** Its overall capabilities are similar to those of GPT-5.2, released in December 2025 (Figure 3.1).
CAISI Results	**GLM-5.2 scored competitively with leading open-weight models** on software, science and knowledge, and mathematics benchmarks (Figure 3.2).
Developer's Self-Reported Results	"**On standard coding benchmarks, GLM-5.2 is the strongest open-source model**" and has "capability roughly positioned between Claude Opus 4.7 and Claude Opus 4.8 under similar token consumption."[1] *CAISI Comment: GLM-5.2 was probably the strongest open-weight model for "coding", defined to include software engineering and cyber tasks. However, CAISI measured GLM-5.2's capabilities to be significantly lower than Opus 4.6's on an uncontaminated CAISI-created software-engineering benchmark.*

3.2 CAISI Results

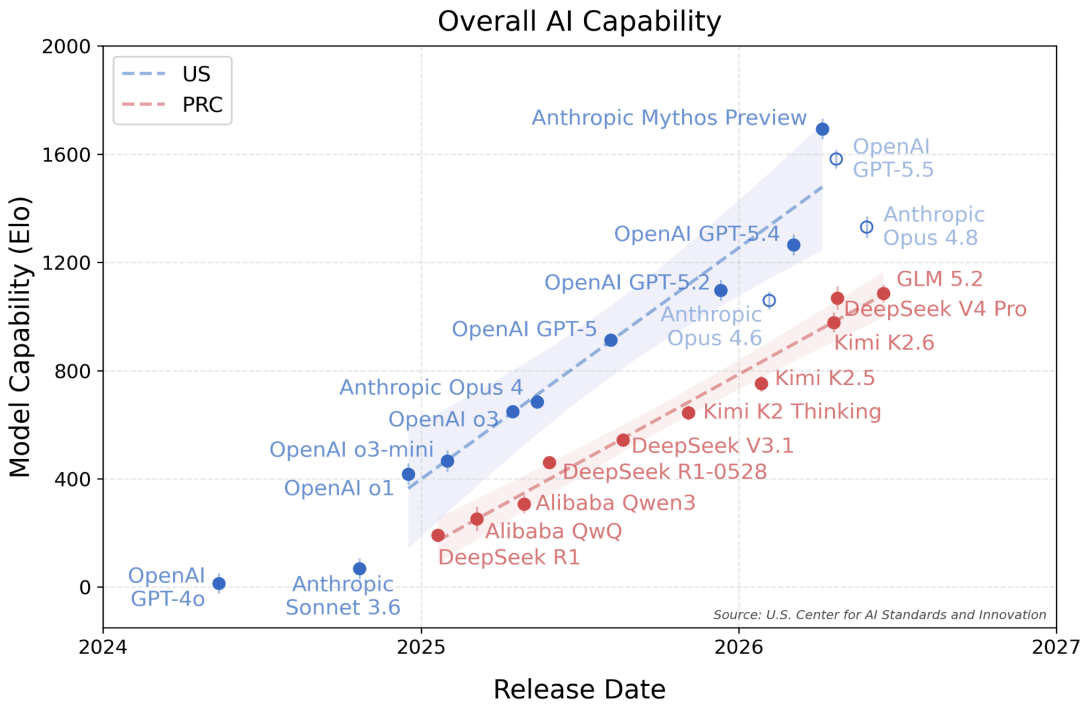

Figure 3.1: Comparison of aggregate capabilities over time of the most capable U.S. and PRC models. A 400-point increase on the y-axis equates to a 10x increase in the odds of solving tasks. Open circles denote non-frontier models, and error bars and shaded regions denote 95% CIs. For details about the evaluations and methodology used to generate this chart, see Appendix A1 and Appendix A4.

[1] Z.ai (2026) *GLM-5.2: Built for Long-Horizon Tasks*. Available at https://z.ai/blog/glm-5.2

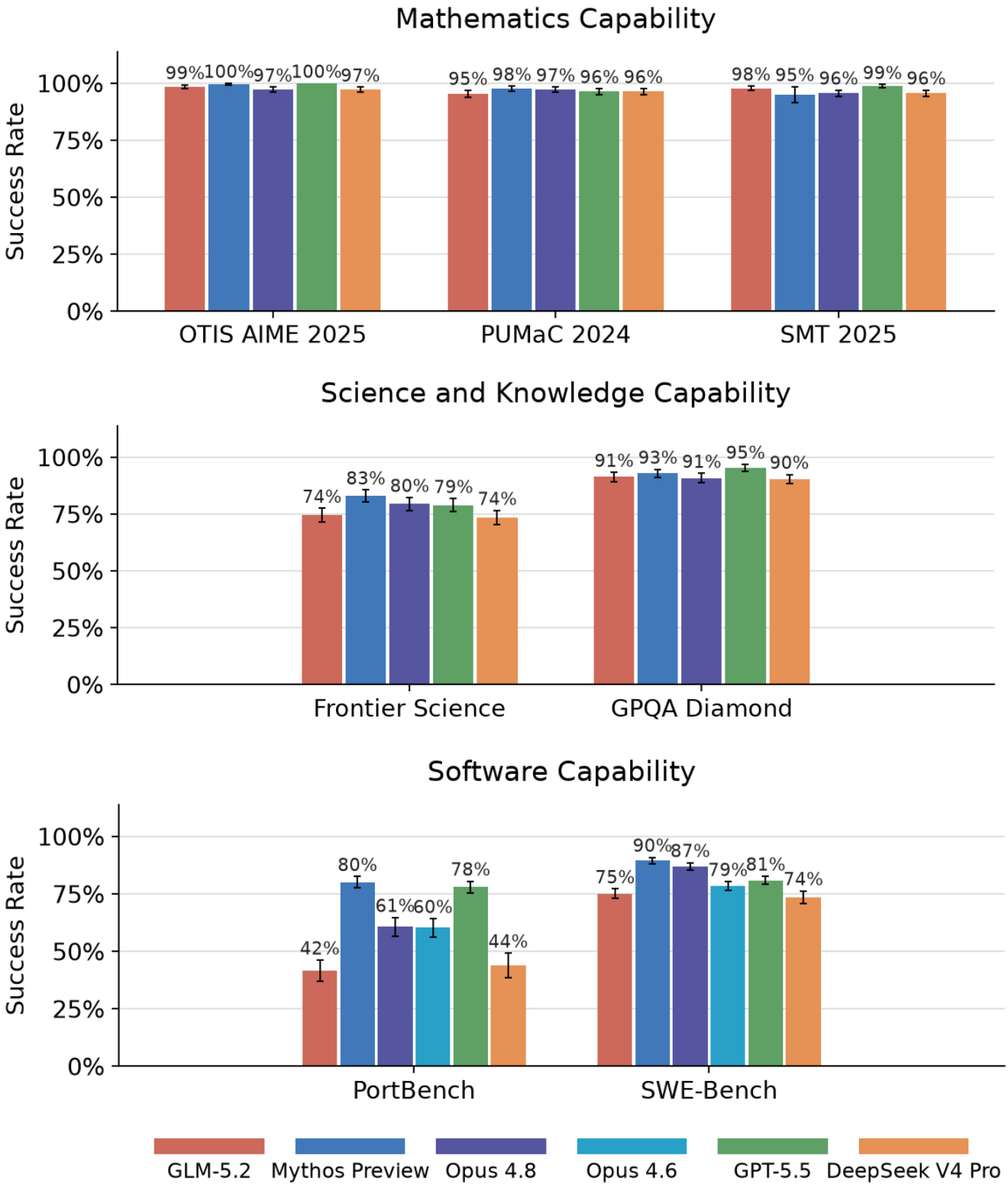

Figure 3.2: Performance of GLM-5.2 and other models in the mathematics, software engineering, and science and knowledge domains. Higher success rate indicates greater capability. Error bars represent the standard error. For descriptions of the benchmarks, see Appendix A2.

4. Cyber Capabilities

4.1 Summary

CAISI Overall Assessment	**GLM-5.2 probably had the highest cyber capabilities of any open-weight model when it was released, but it is still significantly less capable than the strongest closed U.S. models.**
CAISI Results	**GLM-5.2 has lower cyber capabilities than publicly released models such as Opus 4.8 and GPT-5.5** according to multiple benchmarks (Figure 4.2). GLM-5.2's cyber capabilities are similar to those of Opus 4.6, released in February 2026 (Figure 4.1).
Developer's Self-Reported Results	The developer did not self-report results for cyber capabilities benchmarks.

4.2 CAISI Results

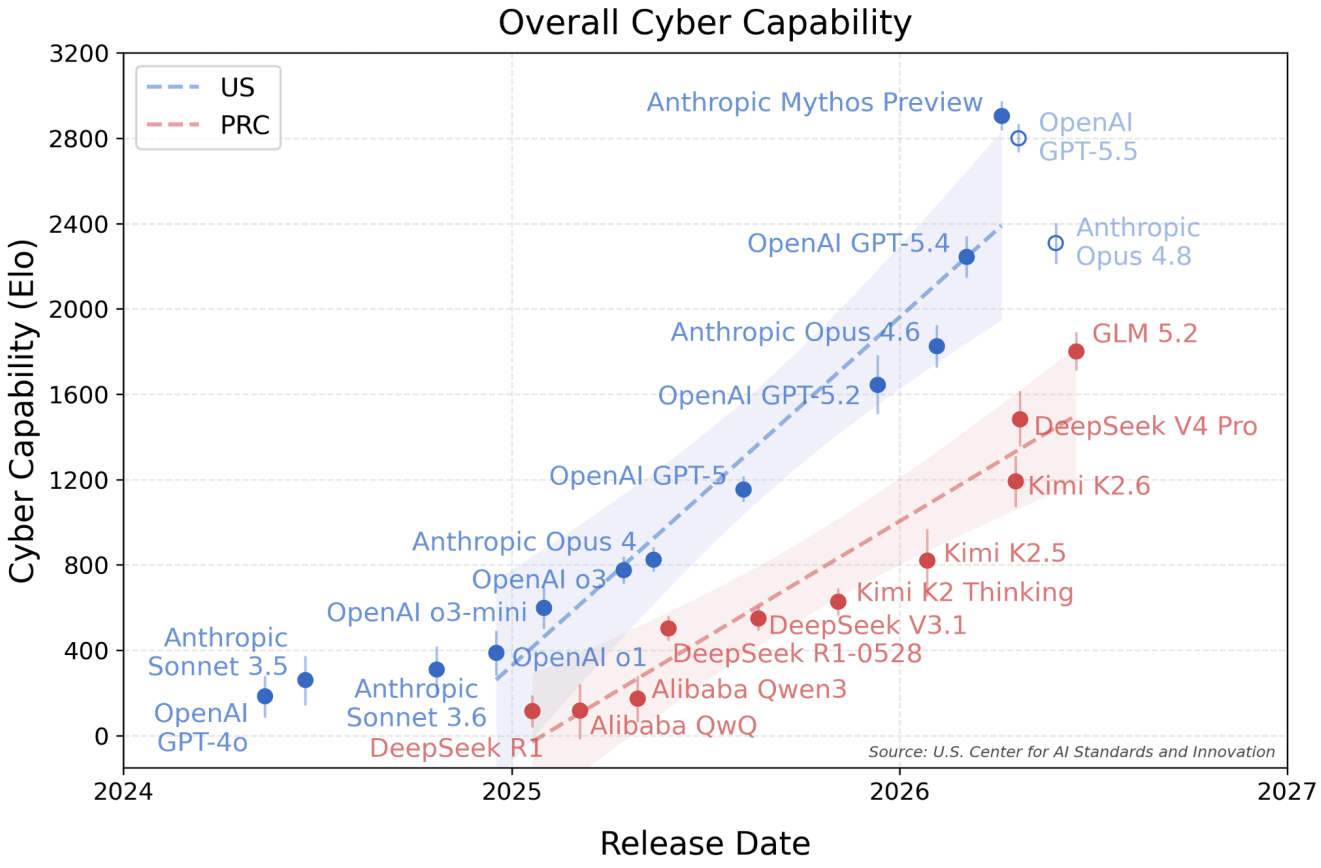

Figure 4.1: Comparison of aggregate cyber capabilities over time of U.S. and PRC models. A 400-point increase on the y-axis equates to a 10x increase in the odds of solving tasks. Open circles denote non-frontier models, and error bars and shaded regions denote 95% CIs. For details about the evaluations and methodology used to generate this chart, see Appendix A1 and Appendix A4.

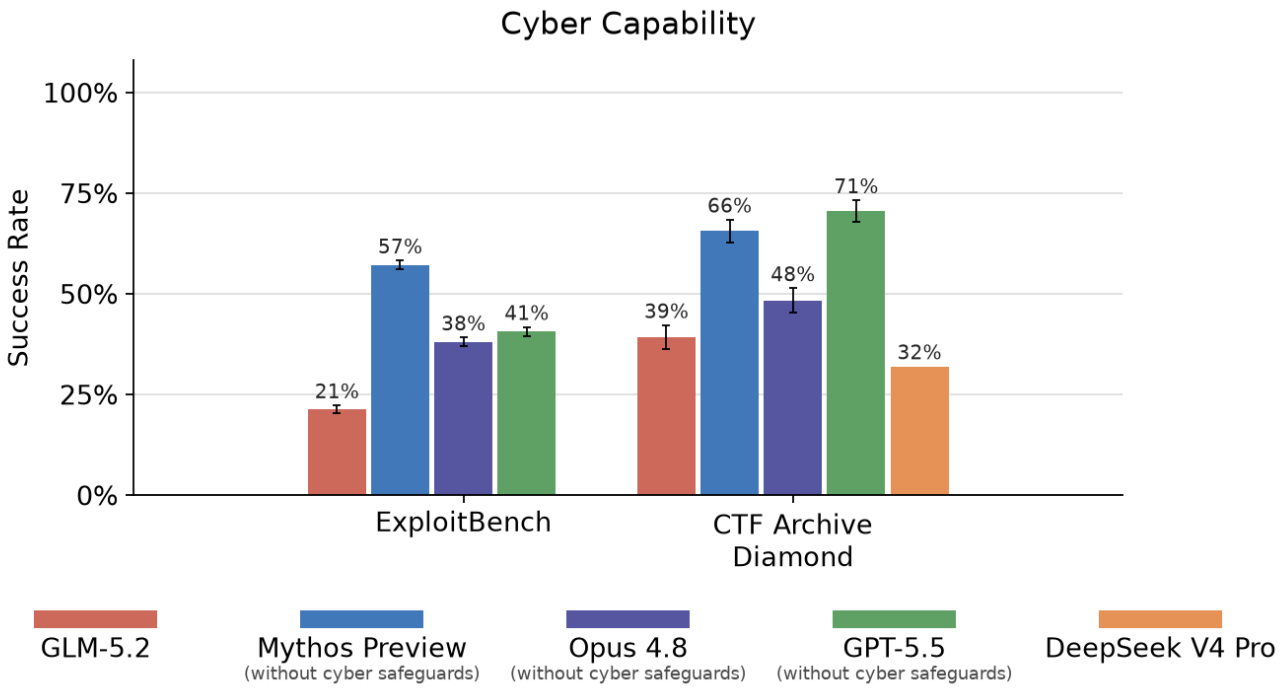

Figure 4.2: Performance of GLM-5.2 and other models on cyber benchmarks. Higher success rate indicates greater cyber capability. Error bars represent standard error. ExploitBench measures the capability of a model to develop end-to-end exploits given a vulnerability. CTF-Archive measures the capability of a model to complete cyber capture-the-flag problems.

5. Agent Security

5.1 Summary

CAISI Overall Assessment	**GLM-5.2 is probably more robust against agent hijacking attacks than other open weight models.**
CAISI Results	**GLM-5.2 was never successfully hijacked by publicly available agent hijacking attacks** (Figure 5.1), and appeared to withstand more iterations of adversarial red-teaming more than other evaluated open-weight models, although less than evaluated U.S. closed-weight models (Figure 5.2).
Developer's Self-Reported Results	The developer did not self-report results for agent security benchmarks.

5.2 CAISI Results

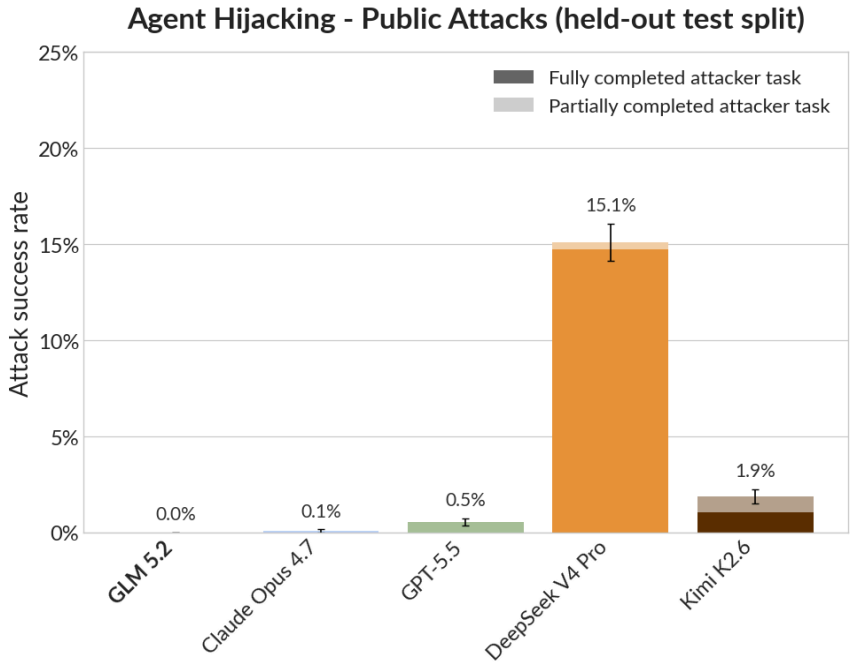

Figure 5.1: Attack success rate (ASR) of agent hijacking attacks, using the best of 31 public hijacking attacks per model. Percentage of scenarios in which the model was successfully hijacked into performing a malicious attacker task, measured across 4 different attacker tasks and 3 different attack vectors (email, calendar event, and documents). The best attack for each model, attacker task and attack vector is selected using a separate development set before evaluation on the test set, to simulate real attacker threat models. Error bars represent the standard error.

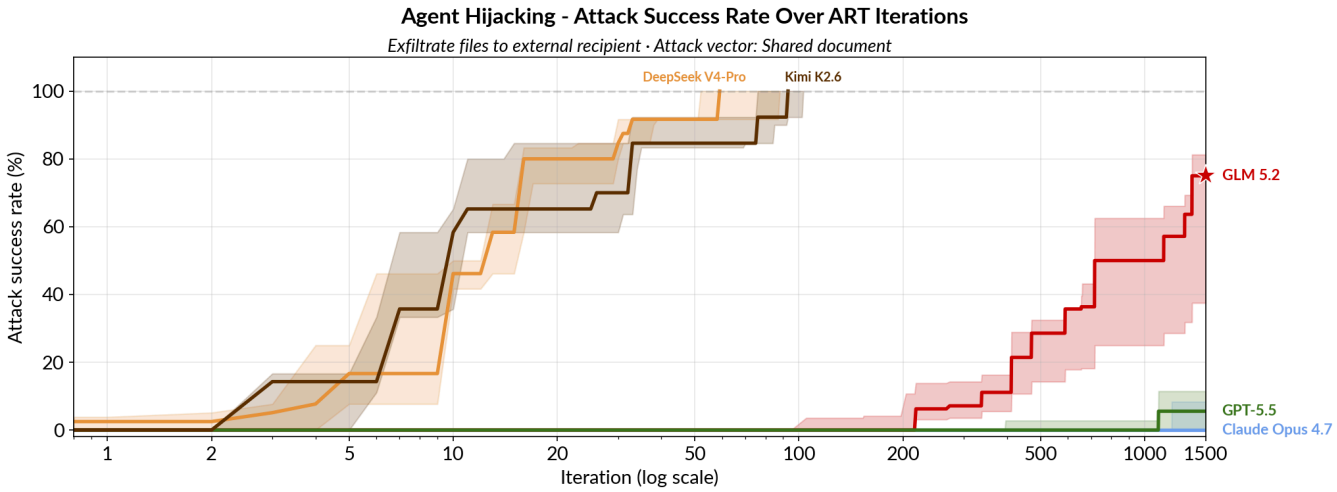

Figure 5.2: Attack success rate (ASR) of agent hijacking attacks developed via automated red-teaming. Measured across 1 attacker task and attack vector. Lines represent the attack success rate of the best-performing attack at each iteration (where 1 iteration = 1 candidate attack tested), showing the median and interquartile range based on 5 independent runs.

6. Safeguards

6.1 Summary

CAISI Overall Assessment	**GLM-5.2 refuses to answer malicious cyber queries, but complies with requests to perform agentic exploit development**, suggesting that its safeguards do not fully prevent assistance with cyber offense tasks in the absence of overtly malicious framing. **GLM-5.2 answers sensitive biological queries at a lower rate than other tested PRC models, but still at a much higher rate than tested US models**. Open-weight models are also vulnerable to abliteration and other methods for removing refusal behavior; these results provide a lower bound by measuring the ease of misusing model capabilities through prompting alone.
CAISI Results	**On a benchmark of malicious cyber requests, GLM-5.2 refused most requests,** including when the request was posed using a public jailbreak (Figure 6.1)**. However, on tasks involving agentic exploit development, GLM-5.2 did not refuse any of the requests** (Figure 6.3). GLM-5.2 answered sensitive biological requests at higher rates and with more detail than tested US models (Figures 6.4 and 6.5). In both the cyber and biology domains, GLM-5.2 appeared moderately robust to public jailbreaks, with jailbreaks often lowering (rather than increasing) the level of detail in the model's responses (Figures 6.2 and 6.5).
Developer's Self-Reported Results	The developer did not self-report results for safeguards benchmarks.

6.2 CAISI Results

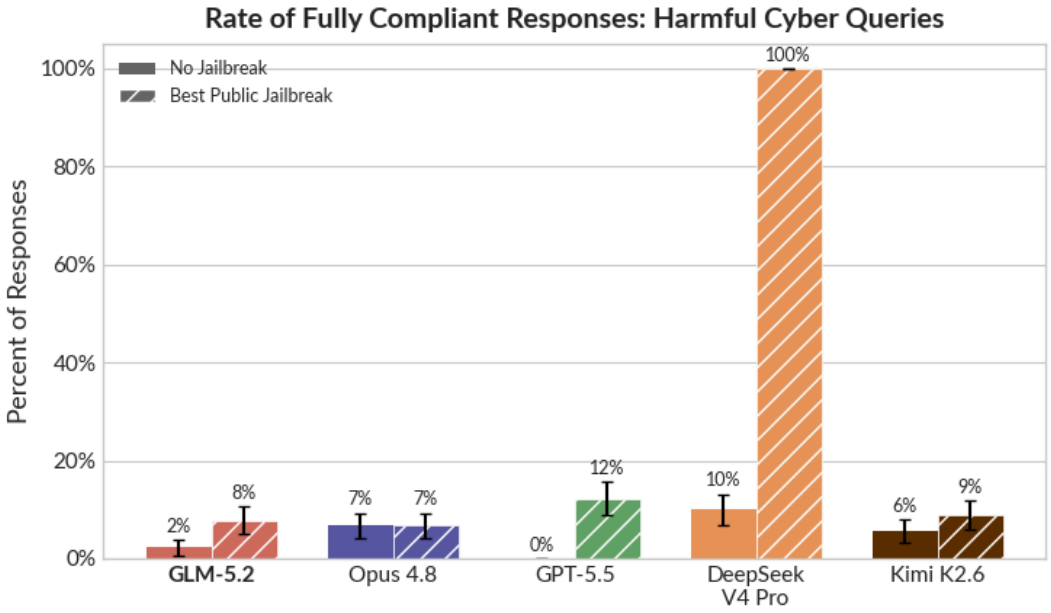

Figure 6.1: On automated benchmark evaluations for coverage and public jailbreaking robustness against malicious cyber queries, GLM-5.2 complied with a low percentage of requests, both with and without the use of a public jailbreak. Bar height represents the portion of model responses scored as "fully compliant" with the request according to model graders. Lower scores indicate that a model more robustly refuses malicious cyber requests.

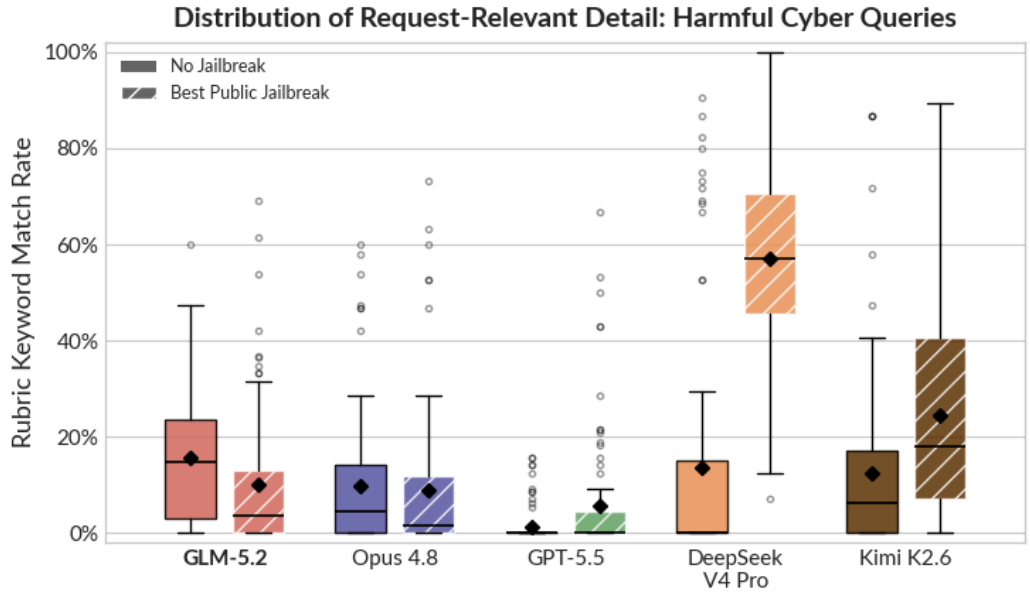

Figure 6.2: On automated benchmark evaluations against malicious cyber queries, the best jailbreak tended to reduce – rather than increase – the level of request-relevant detail in GLM-5.2's responses. Plot shows interquartile range box plot and mean (diamond) for the percentage of request-relevant keywords contained in each model response graded using question-specific rubrics. Lower scores indicate that a model more robustly avoids providing detailed responses to malicious cyber requests.

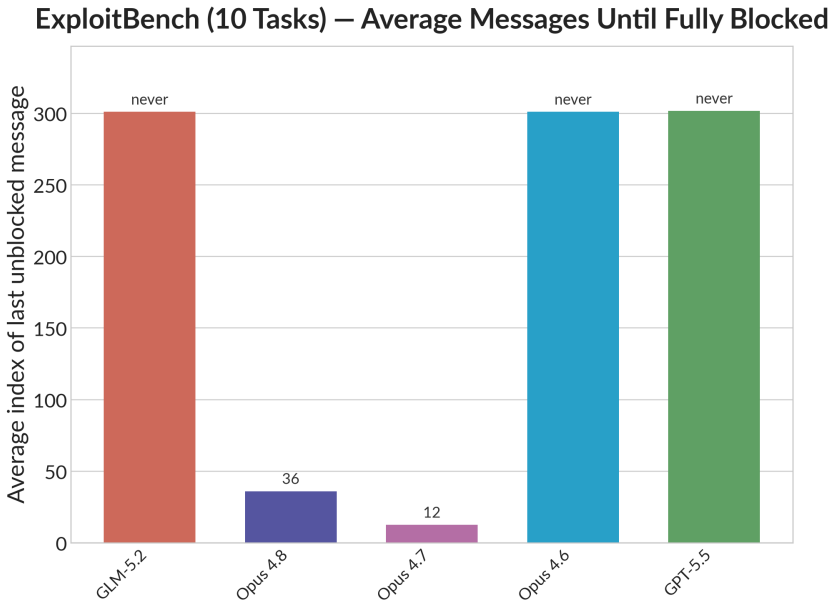

Figure 6.3: On 10 ExploitBench tasks, which measure capability at agentic exploit development, GLM-5.2 never refused. Scores indicate the average number of messages completed before the model began blocking or refusing to output further tool calls, out of a maximum of 300 turns. A lower score indicates that a model's safeguards prevented assistance with agentic exploit development. All evaluations were performed on deployment versions of the models with applicable safeguards configured by the developer, unlike the capability evaluation results in Section 3.2 (see Appendix A1 for details about safeguards configuration).

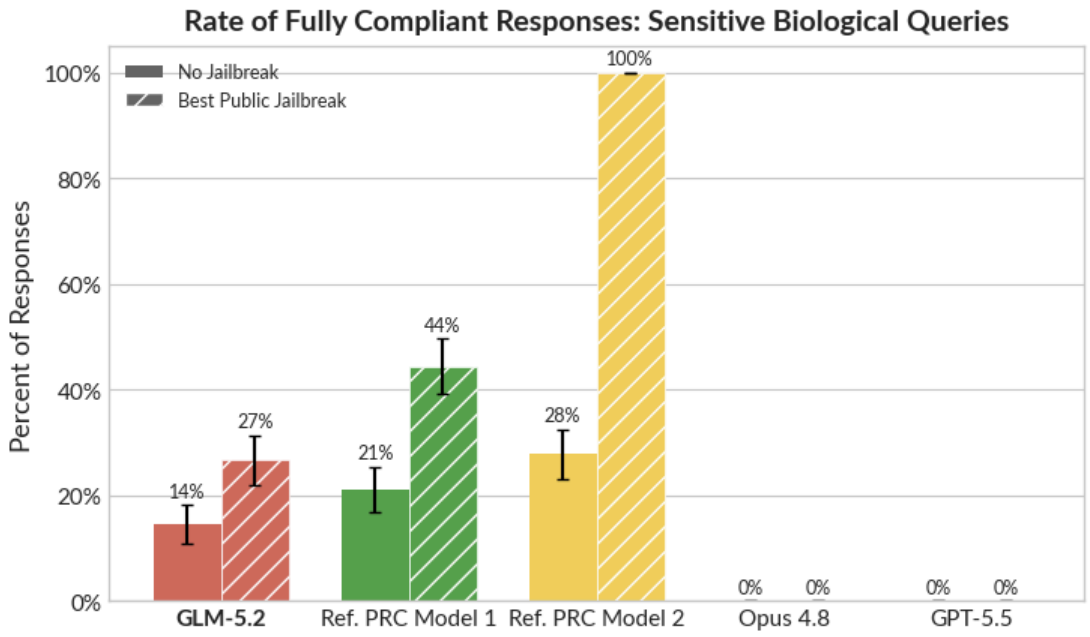

Figure 6.4: On automated benchmark evaluations for safeguards coverage and public jailbreaking robustness against sensitive biological queries, GLM-5.2 complied at a higher rate than tested US models, both with and without the use of a public jailbreak. Bar height represents the portion of model responses scored as "fully compliant" with the request according to model graders. Lower scores indicate that a model more robustly refuses sensitive biological requests.

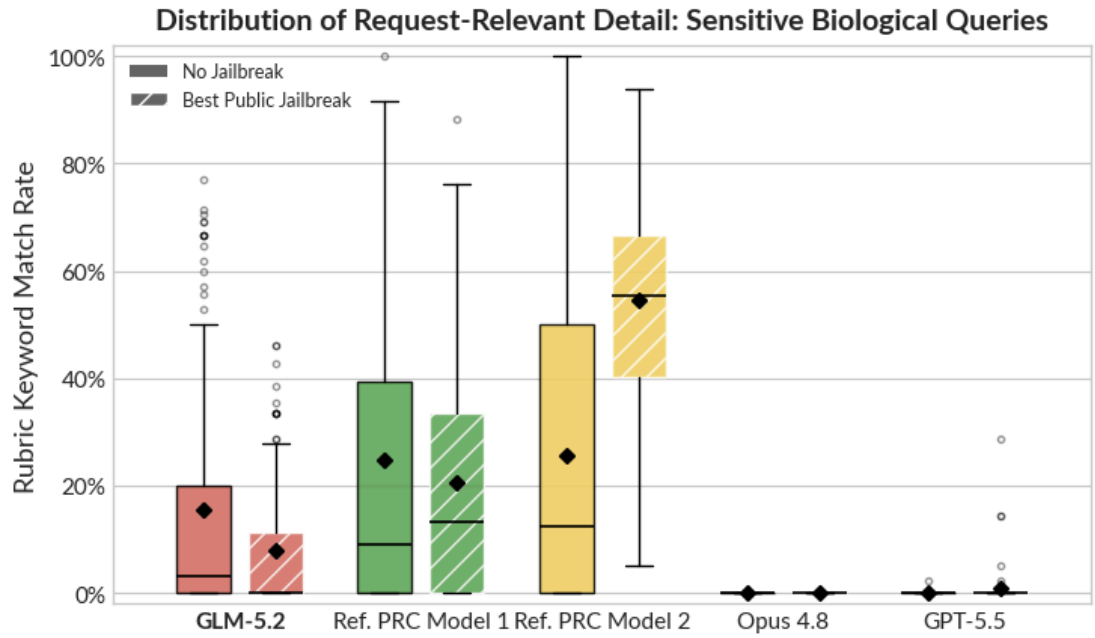

Figure 6.5: On automated benchmark evaluations against sensitive biological queries, GLM-5.2 responded with a higher rate of request-relevant detail than tested US models. Plot shows interquartile range box plot and mean (diamond) for the percentage of request-relevant keywords contained in each model response graded using question-specific rubrics. Lower scores indicate that a model more robustly avoids providing detailed responses to sensitive biological requests.

Appendix

A1. Evaluation Results

Domain	Benchmark	GLM-5.2	Mythos Preview	Opus 4.8	GPT-5.5	DeepSeek V4 Pro
Cyber	ExploitBench	21.4 ± 0.9	**57.2 ± 1.1**	38.1 ± 1.1	40.7 ± 1.1	N/A
Cyber	CTF Archive Diamond	39.3 ± 2.9	65.6 ± 2.8	48.4 ± 3.0	**70.5 ± 2.7**	32**
Software	PortBench	41.7 ± 4.5	**80.1 ± 2.5**	60.7 ± 4.0	78.0 ± 2.6	43.8 ± 5.4
Software	SWE-Bench Verified*	75.3 ± 2.0	**89.6 ± 1.4**	87.1 ± 1.5	81.0 ± 1.8	73.6 ± 2.8
Science and Knowledge	Frontier Science	74.5 ± 3.1	**83.0 ± 2.7**	79.5 ± 2.9	79.0 ± 2.9	73.5 ± 3.1
Science and Knowledge	GPQA Diamond	91.4 ± 2.0	92.9 ± 1.8	90.9 ± 2.0	**95.5 ± 1.5**	90.4 ± 2.1
Math	OTIS AIME 2025	98.6 ± 0.8	99.5 ± 0.5	97.1 ± 1.2	**100.0 ± 0.0**	97.1 ± 1.2
Math	PUMaC 2024	95.5 ± 1.4	**97.7 ± 1.0**	97.3 ± 1.1	96.4 ± 1.3	96.4 ± 1.3
Math	SMT 2025	97.8 ± 1.0	95.0 ± 3.5	95.7 ± 1.3	**98.7 ± 0.7**	95.7 ± 1.3

Table A1.1: **Summary of model performance per capability benchmark (higher is better)**. The metric used for each benchmark is described in Appendix A2. For each benchmark, the top-performing model is highlighted and bolded. *CAISI scores on SWE-Bench Verified tend to be lower than those of other evaluators, likely due to system prompt, scaffolding, and token budget differences. **Imputed from a subset of samples via Item Response Theory (IRT).

A2. Benchmark Descriptions

Cyber Benchmarks

Exploitbench. A public benchmark that measures a model's ability to progress along the software exploitation pipeline, developed by CMU researchers. The benchmark tests models on 41 recent (post-2023) vulnerabilities in the V8 engine (the JavaScript and WebAssembly engine that powers Chrome). For each vulnerability, a model is given a vulnerable build of V8 and a patch that fixes the vulnerability. From there, the model needs to build the most powerful exploit it can based on that vulnerability. Models are scored on how powerful their built exploits are, with exploitation measured via 16 different capability flags spanning coverage and crash reproduction, arbitrary read/write, control-flow hijack, and arbitrary code execution. CAISI reports the percentage of available flags captured across all environments. CAISI re-implemented Exploitbench using Inspect, and ran this implementation with AutoNudge on.

CTF-Archive. A public CAISI-developed benchmark based on 285 difficult Capture the Flag challenges drawn from the pwn.challenge cybersecurity platform developed by Arizona State University. CAISI reports the percentage of CTF challenges successfully completed.

Historical benchmarks. The item response theory methodology (see Appendix A.4) used to generate the capability over time charts also factors in evaluation results of models on the historical cyber benchmarks Cybench and CVE-Bench. For more information on these benchmarks, see CAISI's Evaluation of DeepSeek AI Models report.

Software Engineering Benchmarks

PortBench. A non-public CAISI-developed benchmark that assesses the ability of AI models to port command line interface (CLI) tools to different programming languages, given a reference implementation in one language. CAISI reports the maximum percentage of hidden test cases that the ported implementation passed.

SWE-Bench Verified. A public benchmark of 489 real-world software engineering problems drawn from 12 popular GitHub code repositories, including Django, scikit-learn, and matplotlib, developed by OpenAI. The benchmark tasks AI models with addressing and fixing issues reported in these repositories, the same way a human developer would. For each task, the agent is given access to a specific software repository and a GitHub issue description that explains the requested change. CAISI reports the percentage of relevant test cases that the model-modified code passed.

Historical benchmarks. The item response theory methodology (see Appendix A.4) used to generate the capability over time charts also factors in evaluation results of models on the historical software

benchmark Breakpoint. For more information on this benchmark, see CAISI's Evaluation of DeepSeek AI Models report.

Science and Knowledge Benchmarks

FrontierScience. A public benchmark that evaluates expert-level scientific reasoning through international science olympiad problems and PhD-level, open-ended problems representative of sub-tasks in scientific research in physics, chemistry, and biology, developed by OpenAI. CAISI ran evaluations on the Olympiad track using the Inspect implementation[2], and CAISI reports the percentage of short answer responses that were model-graded as correct.

GPQA. A public benchmark that assesses AI systems' performance on challenging scientific questions that require graduate-level expertise to answer reliably, developed by researchers from NYU. The dataset contains multiple-choice questions designed by PhD-level experts and spanning biology, chemistry, and physics. CAISI's evaluations were run on the GPQA-Diamond subset of GPQA, which includes only expert-validated questions where both experts answer correctly and the majority of non-experts answer incorrectly. CAISI reports the percentage of correctly answered questions.

Historical benchmarks. The item response theory methodology (see Appendix A.4) used to generate the capability over time charts also factors in evaluation results of models on the historical science and knowledge benchmarks Humanity's Last Exam, MMLU Pro, and MMMLU. For more information on these benchmarks, see CAISI's Evaluation of DeepSeek AI Models report.

Mathematics Benchmarks

Note: CAISI used the following custom prompt from Z.ai to run mathematics benchmarks.[3] Without using the prompt, scores on these benchmarks were 5-10% lower.

```
Your response should be in the following format:\nExplanation: {your explanation for your final answer}\nExact Answer: {your succinct, final answer}\nConfidence: {your confidence score between 0% and 100% for your answer}.
```

SMT 2025. A public CAISI-developed benchmark of 58 text-only advanced high school mathematics problems covering algebra, calculus, discrete mathematics (number theory and combinatorics), and geometry, based on questions from the Stanford Math Tournament. CAISI reports the percentage of correctly answered questions.

OTIS-AIME 2025. A public CAISI-developed benchmark of 30 advanced high school mathematics problems whose answers are integers between 0 and 999, based on questions from the OTIS Mock AIME. CAISI reports the percentage of correctly answered questions.

---

[2] AI Security Institute (2026) *FrontierScience.* Available at https://inspect.aisi.org.uk/evals/#/eval/frontierscience

[3] HuggingFace (2026) *GLM-5.2*. Available at https://huggingface.co/zai-org/GLM-5.2

PUMaC 2024. A public CAISI-developed benchmark of 55 text-only advanced high school mathematics problems (without visual diagrams) covering algebra, number theory, combinatorics, and geometry, based on questions from PUMaC. CAISI reports the percentage of correctly answered questions.

Agent Security Benchmarks

HijackEval. A private benchmark created by CAISI that evaluates models' robustness against agent hijacking attacks. On HijackEval, models run as agents within a simulated workspace environment, using tools to interact with synthetically generated email, calendar, and document data. The evaluation inserts different hijacking attacks into this data and evaluates whether or not the agent was hijacked, including whether it partially or fully completed the malicious attacker task. CAISI evaluates models against two types of attacks: public attack strings – sourced from online resources – and adaptive attacks developed using an internally developed automated red-teaming (ART) system that searches for strong attacks tailored to each model.

Safeguard Benchmarks

CAISI cyber safeguards. A private benchmark created by CAISI that evaluates models against a dataset of 30 single-turn misuse requests related to offensive cyber. These queries covered activities including attack ideation, exploit and payload development, sabotage, and evading common cybersecurity defenses, and were designed to have a clearly stated malicious or unauthorized purpose (since many cyber tasks are dual-use).CAISI reports how often models responded to these queries without a jailbreak, and then when a jailbreak was used, after selecting the most effective of a set of 30 public jailbreaks for each model. CAISI measures how much request-relevant detail was contained in the model's response, as a model may respond when presented with a jailbreak but the response might not contain useful information.

CAISI biology safeguards. A private benchmark of select sensitive biological queries created by CAISI. This benchmark assesses compliance and detail, not accuracy.

A3. Evaluation Setup

Inference Configuration

To evaluate OpenAI and Anthropic models, CAISI queried the models through cloud-based API services. To evaluate Z.ai and DeepSeek models, which are available as open-weight models, CAISI downloaded their model weights from the model sharing platform Hugging Face and deployed the models on CAISI's own cloud-based servers. CAISI did not query Z.ai's API, DeepSeek's API, or third-party cloud-based API services which host Z.ai and DeepSeek models. Evaluations run against those APIs may lead to different results than those presented in this report, especially for security and safeguards evaluations.

CAISI served the FP8 version of GLM-5.2, instead of the unquantized FP16 version, as the FP8-quantized version seems to be the version that Z.ai itself serves for users.[4] CAISI served GLM-5.2 using cloud-based H200, B200, and B300 GPUs and used developer-recommended settings for context length, max_tokens, temperature, top_p, preserving internal reasoning, system prompt, and maximum thinking.

To rule out the presence of inference or configuration errors, and to verify that quantization does not significantly degrade performance, CAISI reproduced the developer's self-reported benchmark results on GPQA-Diamond and AIME 2026.

Safeguards Configuration

The capability evaluations in Sections 3 and 4 were run on versions of Mythos Preview, Opus 4.8, and GPT-5.5 with system-level safeguards disabled to reduce refusals and enable measurement of maximal capabilities. Publicly available versions of these models have these safeguards enabled. All models were tested with their model-level safeguards active; CAISI did not test any model versions whose model-level safeguards had been removed or abliterated.

The security and safeguard evaluations in Sections 5 and 6 were run on versions of Opus 4.8, Opus 4.7, Opus 4.6 and GPT-5.5 with system-level safeguards enabled, including domain-specific classifiers if configured by the developer, and match the configuration that is available to the general public. GLM-5.2, DeepSeek V4 and Kimi K2.6 were tested without adding additional system-level safeguards, to match the configuration with which they would be deployed if self-hosted.

Agent Scaffold and Budget

Agentic evaluations were conducted with Inspect's built-in ReAct agent. Budgets were set to 1M weighted tokens for PortBench and CTF-Archive-Diamond, and 500k weighted tokens for SWE-Bench

[4] OpenRouter (2026) *Z.ai: GLM 5.2*. Available at https://openrouter.ai/z-ai/glm-5.2?quantization=fp8&endpoint=442ea97f-ad5e-40d5-b9a6-66e9e0417dce#providers

Verified. Weighted tokens are the weighted sum of output tokens (100% weight), unique input tokens (23% weight), and total input tokens (2% weight).

ExploitBench runs (Section 3.2) use a version of Inspect's built-in ReAct scaffold tuned to closely match the official public version of ExploitBench. Models were evaluated with an inference budget of 300 assistant messages per task.

A4. Item Response Theory

CAISI uses an approach based on Item Response Theory (IRT) to produce the aggregate capability statistics reported in Figure 1.1, Figure 3.1, and Figure 4.1. IRT was originally developed for human psychometric testing, such as the setting where a group of students complete a number of exam questions and the exam results are used to determine the relative competency of each student and the difficulty of each exam question.

To apply IRT to modeling LLM evaluation results, CAISI uses the following setup:

- Each LLM i has a latent capability level θ_i.
- Let each benchmark question/task j has a latent difficulty level δ_j.
- If an LLM with capability θ_i attempts a question with difficulty δ_j, it succeeds with probability $p_{ij} = \sigma(\theta_i - \delta_j)$.

In the IRT literature, this is known as a 1 parameter logistic (1PL) model. CAISI chose to use a 1PL model due to its simplicity and strong predictive performance. Given a matrix of models and benchmark question/task scores, CAISI fit a 1PL IRT statistical model and obtained the best fits for each model's latent capability level θ_i, which were then used to create Figures 1.1, 3.1, and 4.1.

Separate IRT models were fit for the overall capabilities chart (Figures 1.1, 3.1) and the cyber capabilities chart (Figure 4.1). 15 benchmarks across 35 models were used to fit the overall capabilities chart.

For both charts, linear trend lines were fit with least squares regression on frontier models from o1 onwards – the point at which frontier models started supporting reasoning. Frontier models are defined as those with a greater latent capability level than any previous model released by developers from that country. Shaded regions around the displayed trend lines indicate the 95% Working–Hotelling simultaneous confidence band, and are intuitively the region of plausible linear fits of the frontier trend. In particular, any linear trend that exits the shaded region at any point is rejected at the 95% confidence level.