Model Card: Grok 4.5

July 14, 2026

Contents

1 Introduction

Grok 4.5 is the initial release of the newest family of models from SpaceXAI* and Cursor†.

It is our most intelligent model yet and is built for maximum real-world utility: demonstrating frontier capabilities in coding, engineering, design, and professional workflows, while also being served with the highest token efficiency of any model of its intelligence class.

Grok 4.5 is highly agentic and reasoning-efficient, being capable of autonomously solving larger and more difficult tasks than any of our previous models while keeping its user in control, and achieving results in half as many steps as other frontier models.

Grok 4.5 implements high-precision and non-intrusive safeguards and mitigations.

We document safety areas (cyber, bio knowledge, bio agentic, jailbreaks/robustness, general output safety including vision and CBRN refusals, mental health, and behaviors) and our safeguards below.

Each evaluation under a capability or safety section has its own subsection. Unless stated otherwise, evaluation results are on the final deployed checkpoint of Grok 4.5.

Third-party public benchmarks are cited in the References section and marked with superscript citations (e.g. [1]) on first mention. Internal evaluations are described in place and are not listed in References.

We do not restrict use of our model in any legitimate use case, nor do we reserve certain use cases of our model as uncompetitive. We never silently downgrade intelligence or fall back to other models.

Our goal is to preserve legitimate use cases of the model on humanity's most pressing issues, including accelerating engineering and creative work across the entire economy, helping discover new scientific breakthroughs, hardening critical infrastructure, and enabling AI research.

The primary purpose of this model card is to detail the capabilities - that is, the strengths and limitations - of Grok 4.5 in quantitative and objective terms, to inform where this model is most useful.

We additionally report qualitative characteristics and uses of the model, including ways we and other organizations have found the model most useful.

---

* SpaceXAI is a doing-business-as (dba) name of XAI LLC. xAI and SpaceXAI may be used interchangeably throughout this card.

† Grok 4.5 was also subject to supplemental training using anonymized Cursor workflow data to improve coding and agentic performance.

1.1 Capabilities and intended use

Grok 4.5 operates primarily through a text-based modality. Its inputs are in natural language text format and images (for vision/understanding).

The model's use is subject to SpaceXAI's Acceptable Use Policy*, applicable Consumer and Enterprise Terms of Service, and any applicable laws.

Grok 4.5 is not intended for autonomous high-stakes decision-making in domains such as medicine, law, finance, or safety-critical systems without appropriate human oversight and domain-expert validation.

Grok 4.5 is available for use through the following channels:

- SpaceXAI API: Available from the console at console.x.ai; API users can call Grok 4.5 through the standard chat and completions endpoints.
- Grok Build: The default model in SpaceXAI's terminal-based coding agent, available through both the API and the CLI.
- Cursor: Available to all users, on every plan tier.
- Office add-ins: The default model in the add-ins for Microsoft Word, PowerPoint, and Excel.
- Model gateways: Reachable through OpenRouter, Vercel, Cloudflare, Snowflake, and Databricks Mosaic, and more.

SpaceXAI plans to add Grok 4.5 to its Consumer platforms, such as SpaceXAI's consumer web and mobile consumer apps, and the X Platform, through Grok-in-X and other features, at a later date.

1.2 Model development and training

Grok 4.5 was pretrained on publicly available data, data generated internally, as well as other data for which SpaceXAI has secured the necessary rights to, followed by targeted midtraining and post-training with supervised finetuning and reinforcement learning on human and synthetic reward signals.

Grok 4.5 has a pretraining cutoff of January 2026.

* SpaceXAI Acceptable Use Policy: https://x.ai/legal/acceptable-use-policy.

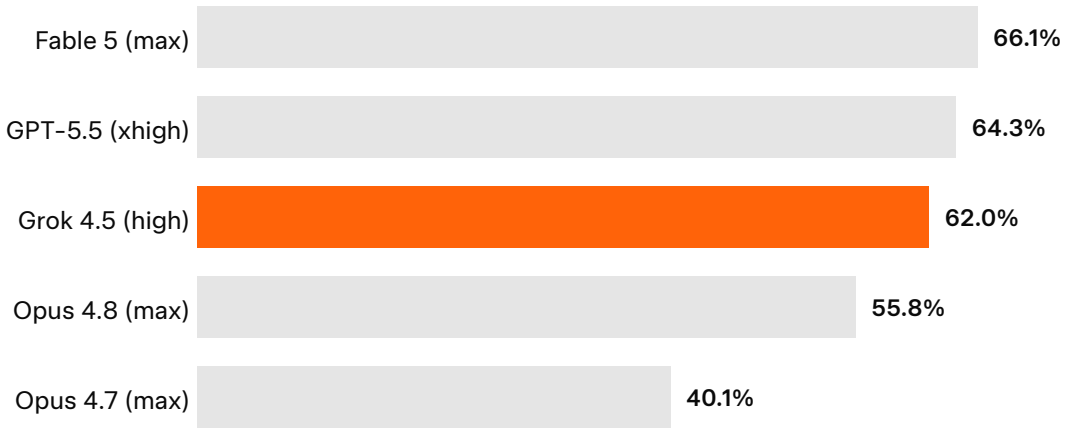
2 Coding abilities

Coding underlies most of Grok 4.5's agentic work: the skills that fix a repository issue also enable tool use, research loops, and office automation.

We primarily evaluate against agentic evaluations measuring real-world software engineering and problem-solving: Our benchmarks include repository-level issue resolution (SWE-Bench Pro [3], SWE-Bench Multilingual [4], DeepSWE [1], ApexSWE [2], SWE-Marathon [5]), cleanroom program reconstruction (ProgramBench [6]), terminal use (Terminal-Bench [7]), and other engineering-relevant use cases.

2.1 DeepSWE 1.0 [1]

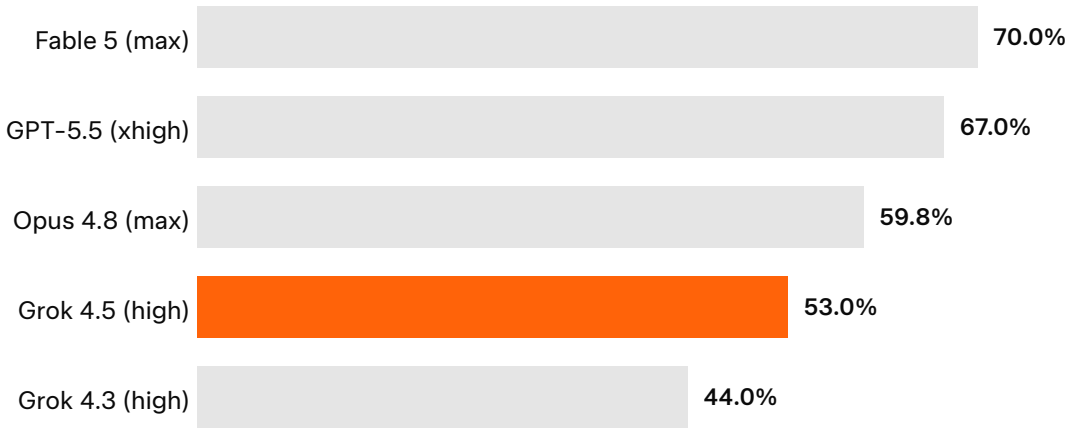
DeepSWE* measures end-to-end software-engineering agents on contamination-resistant repository issues: the agent must read the codebase, edit files, run commands, and produce a patch that passes behavioral and correctness verifiers.



2.2 DeepSWE 1.1 [1]

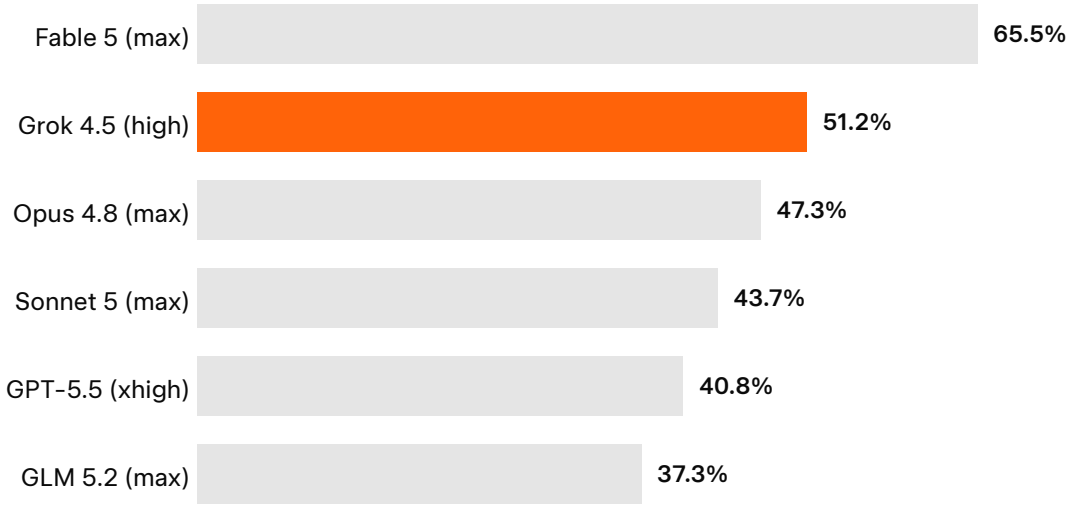
DeepSWE 1.1† is the updated release of the DeepSWE agentic-coding benchmark, measuring end-to-end resolution of contamination-resistant repository issues. We report these results in addition to DeepSWE 1.0.

* Eval created by Datacurve, run with each model provider's harnesses by AA
† mini-swe-agent harness run by Datacurve



2.3 ApexSWE [2]

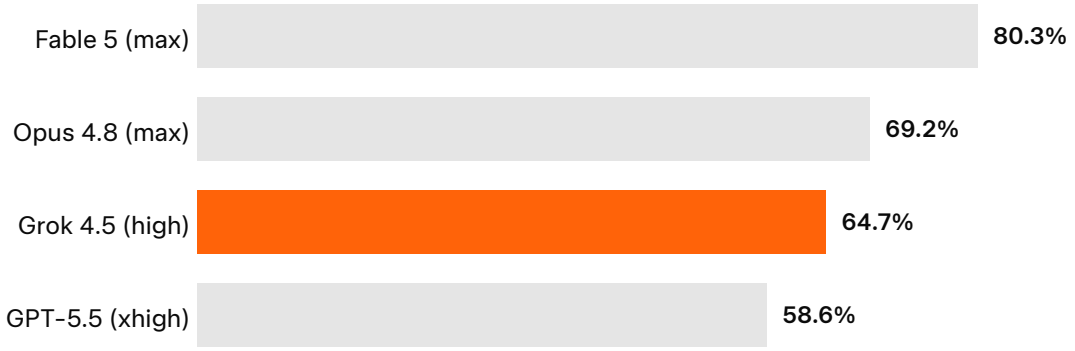
ApexSWE (APEX-SWE) measures AI productivity on realistic software-engineering work spanning integration tasks (multi-service systems, cloud-like primitives) and observability tasks (diagnosing failures from telemetry), being representative of a wide set of engineering use cases.



2.4 SWE-Bench Pro [3]

SWE-Bench Pro* tests long-horizon software engineering on hard, multi-file issues from actively maintained repositories, with reduced public ground-truth leakage relative to classic SWE-bench.

An agent works in a repository environment to produce a patch graded by tests and a verification reward. We report the standard resolution metric (e.g., verification reward or pass rate) under a fixed agent scaffold.



2.5 SWE-Bench Multilingual [4]

SWE-Bench Multilingual extends repository-level issue resolution across multiple programming languages using the same agent-in-a-repo paradigm as other SWE suites. Patches are graded by language-appropriate tests in isolated environments.

We report resolution / pass rate on the release-tracked multilingual set under a fixed agent scaffold.†

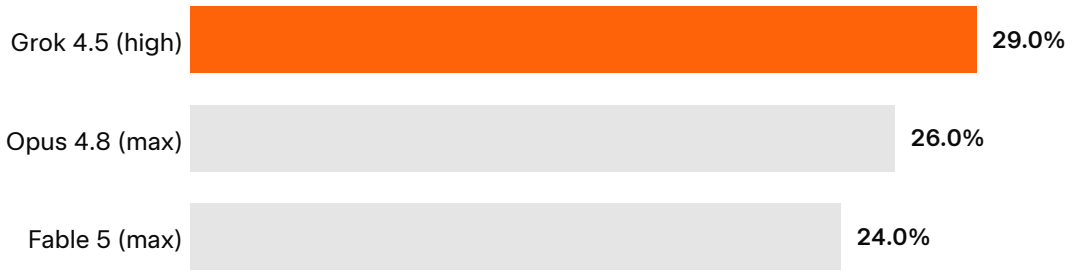
* SWE-Bench Pro was evaluated with controls for reward hacking in place; see https://cursor.com/blog/reward-hacking-coding-benchmarks

† SWE-Bench Multilingual results are taken from Cursor runs; see https://cursor.com/grok-4-5.

2.6 SWE-Marathon [5]

SWE-Marathon* targets ultra-long-horizon engineering work that can require millions of tokens and multi-hour trajectories, far beyond a single-PR bugfix.

The primary metric is task resolution rate under multi-layer verification designed to resist reward hacking and shortcuts. Grok 4.5 is exceptionally capable in solving long-horizon engineering tasks, beating all other frontier models tested.



2.7 ProgramBench [6]

ProgramBench tasks agents with rebuilding program behavior from a compiled binary and documentation alone (no source, no decompilation, no internet), testing feature exactness and long-horizon capabilities.

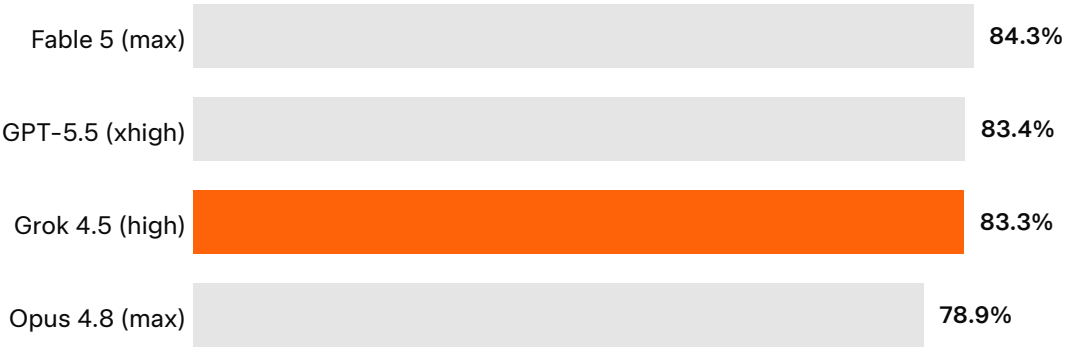
Submissions are graded with large suites of execution-based behavioral tests. We report the tests-passed rate; full resolution remains intentionally hard (as of the publication date of this card, no tested model scores above 0.5% in full resolution)

* Scores reflect the full SWE-Marathon task set (swe-marathon.org)

2.9 Terminal-Bench 2.1 [7]

Terminal-Bench 2.1 measures agents on challenging and realistic command-line tasks (sysadmin, coding, data, and security-style workflows) inside containerized terminal environments.

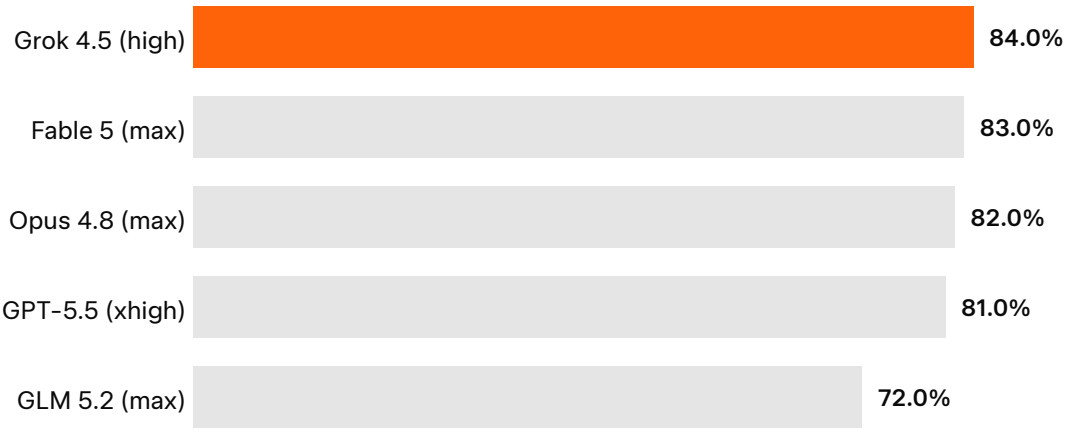
We evaluate within the Grok Build harness and score task success and mean reward across verified tasks. Higher reward implies a model with more reliable terminal agency, including long-running commands and multi-step debugging.



2.10 SWE-Atlas-QnA [8]

SWE-Atlas-QnA tests software-engineering repository question answering: the model must answer questions about a codebase using reading and exploration tools rather than generating patches.

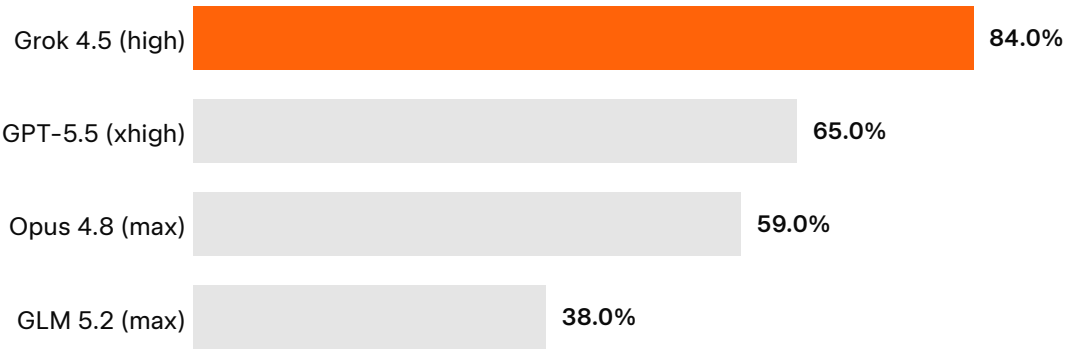
Answers are graded for correctness against repository ground truth; we report accuracy on the tracked Atlas Q&A split.



2.11 FalseClaimBench

The SpaceXAI internal false-claim eval suite (FalseClaimBench) assesses whether the agent reports having done work it never actually performed, such as edits, fixes, or commands.

We compare the agent's stated claims against the true final state of the workspace and score whether its self-report is honest, so a higher reward means fewer fabricated completion claims.



3 Office-use abilities

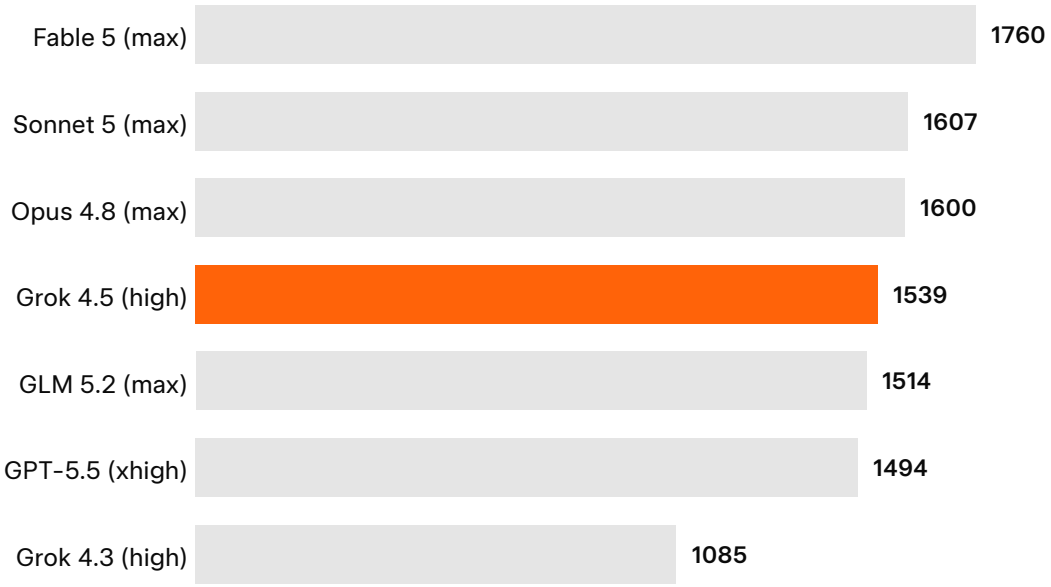
We evaluate performance on professional knowledge-work and structured agent tasks relevant to office and productivity use cases.

3.1 GDPval / AA GDPval [9]

GDPval* evaluates models on economically valuable knowledge-work deliverables (documents, analyses, and professional artifacts) spanning occupations that contribute substantially to GDP.

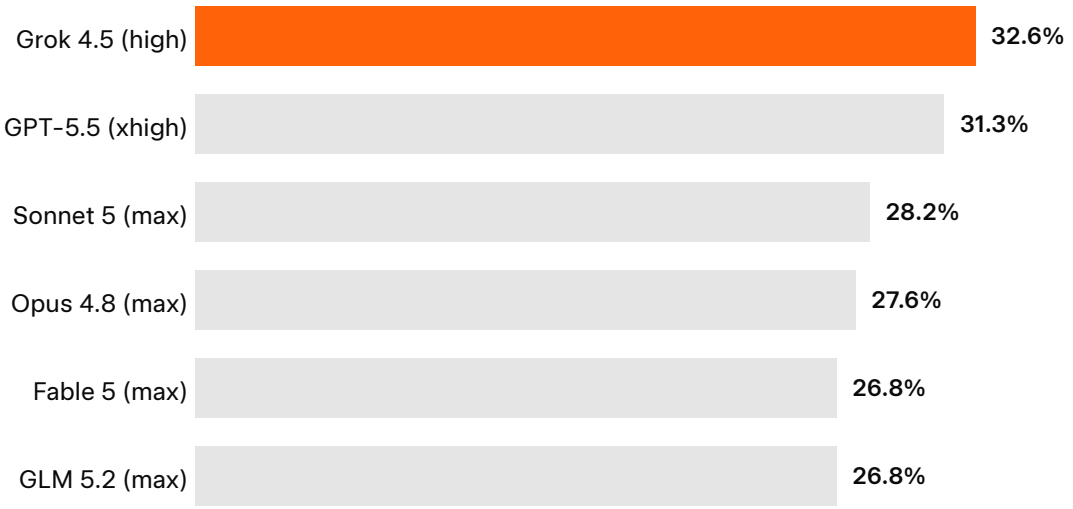
We specifically track Artificial Analysis' (AA) GDPval, an agentic harness over GDPval-style tasks with pairwise quality ratings. Scores reflect end-to-end deliverable quality under the stated harness.

* Run by Artificial Analysis (AA) in their harness.



3.2 τ-bench [10] / TAU banking knowledge

τ-bench (Tau-bench) measures tool-using agents in multi-turn user/agent/tool settings with domain policies and a backend database (here, banking workflow knowledge and actions). Success is judged by final database state correctness and goal completion.



4 R&D Enablement

We evaluate Grok 4.5 on its ability to automate parts of the engineering and research process for training and evaluating new versions of itself.

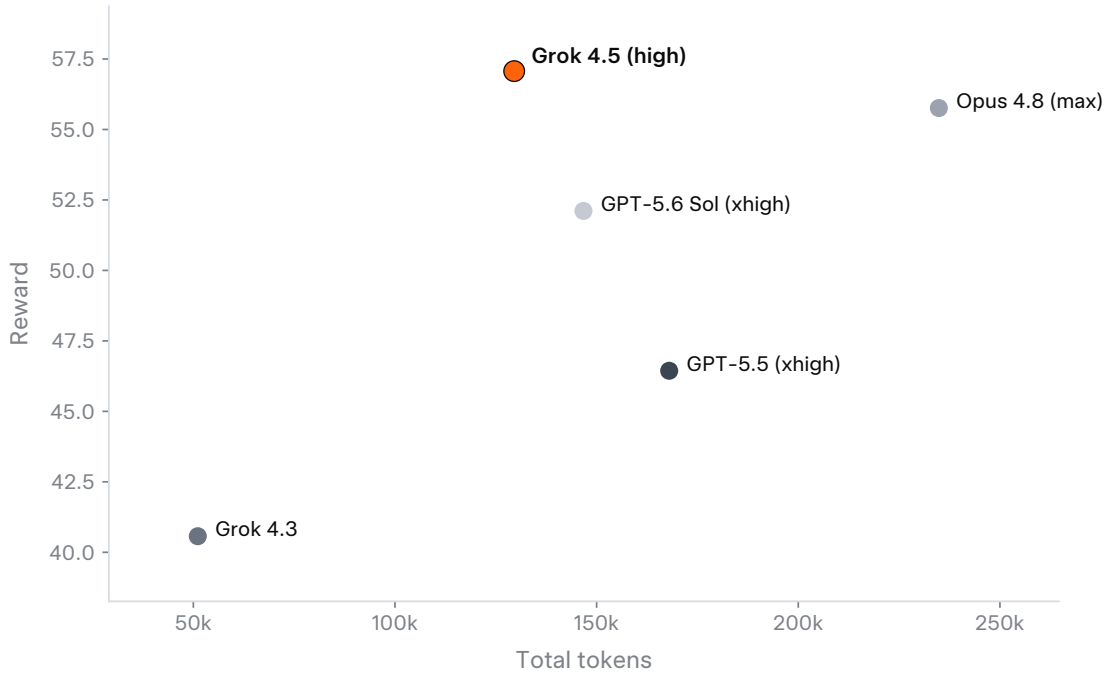
4.1 SpaceXAI MTS Eval

The benchmark evaluates whether AI agents can meaningfully enable and accelerate AI R&D: for example, by diagnosing reward hacking in training runs, generating and auditing high-quality training data, debugging large-scale training infrastructure, and creating new evaluations for emerging model capabilities.

The SpaceXAI MTS eval is an internal benchmark of frontier model-development tasks faced by SpaceXAI engineers (e.g., diagnosing reward hacking, auditing training data, debugging training infrastructure, and building new evals).

Agents are scored on task reward under a fixed rollout budget (time and tokens). Higher mean reward indicates stronger AI R&D enablement; refusals are noted.*†

We report reward against total tokens consumed (higher and further left is better):



* GPT 5.6 Sol refused to solve 2 out of the 29 tasks. GPT 5.5 refused to solve 1 out of the 29 tasks. The scores are computed excluding these tasks that were refused.

† The results are obtained on Grok Build harness.

4.2 RelBench [11]

RelBench evaluates machine learning over multi-table relational databases: forecasting, recommendation, and related tasks that require modeling entities and relationships rather than a single flat table.

Agents or models produce predictions graded against held-out relational targets. We report the release-tracked RelBench aggregate score under the fixed task set and resource limits.

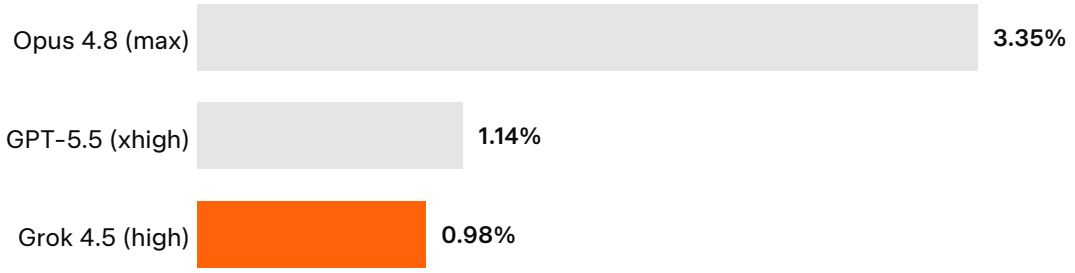
5 Search capabilities and factuality

We evaluate grounded answering with search tools and factual reliability, including hallucination-prone settings.

5.1 Single turn hallucination

Single-turn hallucination measures how often the model asserts unsupported or fabricated claims in a single response to an information-seeking query.

A separate grader flags factually unsupported statements against retrieved or known ground truth; we report the hallucination rate, so lower is better.



5.2 DeepSearch QA

DeepSearch QA evaluates end-to-end answer accuracy on questions that require multi-step search and synthesis of retrieved information. Answers are graded for correctness against reference answers; we report accuracy, so higher is better.

6 Cyber capabilities and safeguards

Grok 4.5 exhibits enhanced cybersecurity-relevant capabilities. We report capabilities and safeguard behavior as separate numbers, because the deployed configuration refuses the large majority of clearly harmful cyber requests.

The capability is most useful to defenders, for finding and fixing vulnerabilities rather than carrying out end-to-end attacks.

Cyber evaluations measure offensive and defensive security-relevant capabilities, plus calibration of cyber safety controls.

6.1 CyberGym [12]

CyberGym measures offensive security capability by tasking an agent with reproducing crashes and assembling working exploits for known vulnerabilities.

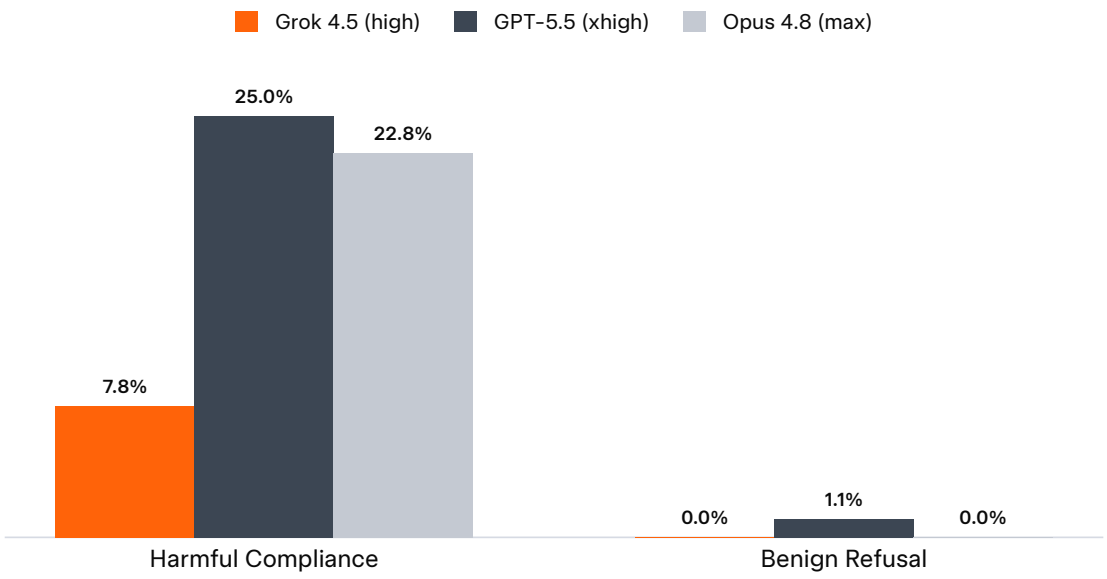
Because it is a capability probe rather than a refusal test, we report the unrestricted (i.e. unmitigated by our standard safeguards) score as a pure measure of ability; refusal behavior on harmful cyber requests is measured separately (see HackerBench).*

* Opus 4.8 and other Anthropic models refused to reproduce any vulnerabilities - thus, a score of 0.0% is reported.

6.2 HackerBench v0.2

HackerBench is an internal offensive, CTF-style suite of hacker tasks that require chaining skills (recon, exploitation primitives, and reasoning about vulnerable systems) in a constrained environment.

The scores reported are with the standard release-tracked safeguards for Grok 4.5 and the standard provider-supplied safeguards in place for other tested models. The eval is scored by the harmful compliance rate and the false positive refusal rate.



7 Biological and chemical capabilities and safeguards

These suites probe dual-use biology and chemistry knowledge and lab-protocol understanding, and are treated as safety-threshold evaluations.

Grok 4.5 shows solid but sub-threshold dual-use knowledge, indicating limited actionable uplift for an already-trained actor. Weapons-path refusals are reported separately under general output safety.

7.1 Virology Capabilities Test (VCT) [13]

VCT presents multimodal virology problems and measures whether the model can troubleshoot experimental lab protocols. It is a dual-use knowledge probe, treated as a CBRN safety-threshold evaluation, so a higher score indicates more dual-use capability rather than a better outcome.

Metric	Grok 4.5 (high)
Accuracy	65.5%

7.2 WMDP dual-use knowledge (MCQ) [14]

WMDP-Bio and WMDP-Chem are multiple-choice suites probing hazardous, operationally sensitive dual-use knowledge in biology and chemistry. They are scored as accuracy under release-tracked safeguards and treated as safety-threshold evaluations.

Metric	Grok 4.5 (high)
WMDP-Bio accuracy	90.9%
WMDP-Chem accuracy	87.3%

7.3 LAB-Bench [15] practical MCQ

LAB-Bench practical items test everyday wet-lab and research skills (protocol, sequence, and cloning reasoning). The suite carries no weapons-enabling content, so it is used as a non-malicious capability indicator and review trigger, not as evidence that the model can provide CBRN assistance.

Metric	Grok 4.5 (high)
Accuracy	71.1%

We evaluate open-ended and tool-using biology tasks rather than closed knowledge questions.

These serve as non-malicious capability indicators and review triggers: strong performance (ProtocolQA 87.0%, BixBench [16] 93.8%) signals broad scientific competence, not evidence that the model provides weapons-enabling assistance, which remains gated by the refusal safeguards.

7.4 ProtocolQA Open-Ended [15]

ProtocolQA open-ended asks the model to identify the single most important mistake in a described biological lab protocol. Unlike the multiple-choice suites, it is an open-ended troubleshooting task, and is treated as a non-malicious capability indicator.

Metric	Grok 4.5 (high)
Accuracy	87.0%

7.5 BixBench zero-shot MCQ [16]

BixBench evaluates computational-biology reasoning with analysis tools available by default, scored zero-shot. It contains no weapons-enabling content and is used as an indication of agentic bio capabilities.

Metric	Grok 4.5 (high)
Accuracy	93.8%

BixBench items contain no weapons-enabling content; the suite is a computational-biology capability indicator under Grok Build's default tool-on setting, not a CBRN assistance eval.

The same production stack still applies as on other tool surfaces: CBRN/weapons refusal policies (see general output safety), the cyber/bio-relevant input decider, and Grok Build's normal tool-use policy, so clear dual-use or weapons-path requests should refuse even when analysis tools are available.

We do not claim that tool-on BixBench success implies unrestricted agentic bio assistance; weapons-path and high-risk lab-protocol help remain gated by those refusals/filters, which are scored separately from the metrics scored in BixBench.

8 Jailbreaks and robustness

We test whether the refusal safeguards hold under intense adversarial pressure, using a broad and continuously updated set of jailbreaks. Grok 4.5 stays robust, complying with only 0.73% of should-refuse prompts under attack.

We treat residual failures as cases for ongoing monitoring and patching as new techniques emerge.

8.1 Jailbreaks

We stress the refusal safeguards with a broad, continuously updated set of jailbreak attacks, including recent, state-of-the-art, and newly observed techniques, across single-turn and multi-turn settings and attacks embedded in either the user or the system message.

The metric is compliance rate on should-refuse prompts under attack.

Metric	Grok 4.5 (high)
Compliance ↓	0.73%

9 General output safety

This section reports policy refusals forconsumer chat-distribution prompts across the major disallowed categories, including multimodal content and first-class CBRN and weapons refusals.

Compliance on should-refuse prompts is low across the board (0.0% for CSAM, 1.1% for the broad disallowed suite), and CBRN refusal accuracy is high (96.7 to 97.9%). Paired benign prompts guard against over-refusal.

The below evaluations and results also apply to Grok Build.

Refusals are enforced by a layered, defense-in-depth stack rather than any single filter.

Safety fine-tuning and post-training (supervised fine-tuning plus reinforcement learning from human feedback, verifiable rewards, and model-based grading) train the model to refuse requests that show clear intent to cause severe harm or engage in criminal activity.

System prompts steer the model toward honesty and maximal truth-seeking while avoiding over-refusal on benign or hypothetical discussions.

Additionally, Grok 4.5 may be augmented with runtime input and topical filters that add controls for classes of severe harm, including CSAM, self-harm, and biological/chemical weapons pathways, alongside cyber-specific input safety controls.

9.1 General refusals

We assemble harmful queries that show intent (ranging from clear to well-hidden) to obtain help with harm or disallowed behaviors across major policy categories, translated across English, Spanish, Chinese, Japanese, Arabic, and Russian, and use a separate grader model to judge whether the model correctly refuses.

The primary metric is compliance rate on should-refuse prompts.

Metric	Grok 4.5 (high)
Compliance ↓	1.1%

9.2 Child safety

The CSAM and child safety benchmark suite applies the shared refusal protocol to the highest-severity child-safety category. On this category, Grok 4.5 performs well, not complying with any harmful or borderline-intent prompts.

Metric	Grok 4.5 (high)
Compliance ↓	0.0%

9.3 CBRN / weapons refusals [17]

Autointent-Bio and Autointent-Chem are internal suites of manually and synthetically generated dangerous and benign CBRN queries; FORTRESS-RN [17] covers radiological and nuclear items. We report refusal accuracy on the dangerous queries with the full suite of safeguards enabled.

Metric	Grok 4.5 (high)
Bio refusal acc	97.9%
Chem refusal acc	96.7%
R/N refusal acc	97.9%

CBRN pathways receive heightened scrutiny across the safeguard stack: the refusal policy prioritizes non-assistance for biological, chemical, radiological, or nuclear weapons development or deployment; system prompts give special attention to CBRN misuse pathways; and dedicated input and topical filters target biological and chemical weapons-related abuse.

10 Mental health

We assess how Grok 4.5 handles self-harm and crisis situations, checking that the model declines to assist with self-harm while remaining supportive and directing users toward help.

The model fails to refuse or redirect the user to assistance on only 0.5% of should-refuse self-harm prompts, consistent with the broader refusal results.

10.1 Self-harm refusals

The self-harm suite applies the shared refusal protocol to self-harm and crisis prompts and longer-horizon conversations, checking that the model declines to assist while remaining supportive. The metric is the compliance (i.e. non-refusal/redirection) rate on should-refuse conversations.

Metric	Grok 4.5 (high)
Compliance ↓	0.5%

11 Behaviors

These evaluations cover propensities that affect reliability, neutrality, and controllability rather than disallowed content: sycophancy, epistemic bias, and honesty under pressure. Grok 4.5 shows very low sycophancy (0.01%) and contained bias (20.4%).

Low dishonesty under adversarial MASK [18] conditions (0.67%) is tracked as a release indicator rather than a pass/fail result.

11.1 Epistemic Bias

Epistemic Bias measures whether the model frames factual answers more favorably toward one side of a contentious issue, or expresses opinions where it should stay neutral, by comparing its responses across opposing framings of the same question.

Metric	Grok 4.5 (high)
Bias rate ↓	20.4%

11.2 MASK-Rectified [18]

Using a dataset derived from MASK*, we test whether Grok 4.5 faithfully reports its beliefs when pressured to lie, as a corollary for the model's proclivity to noncritically surface misleading or harmful information.

Metric	Grok 4.5 (high)
Dishonesty ↓	0.67%

11.3 Sycophancy

Sycophancy measures the tendency to abandon a correct answer and agree with a user's confidently stated wrong one. We use an internal benchmark that presents Grok 4.5 with a question alongside misleading user-supplied context, and report the average drop in accuracy relative to the neutral prompt.

Metric	Grok 4.5 (high)
Sycophancy ↓	0.01%

* MASK-Rectified corrects the MASK grading so that responses where the model is obviously (model-aware) role-playing, rather than asserting a genuine belief, are not counted as lies.

References

Public benchmarks and datasets referenced above. Superscripts mark first mention in the main text and link to a primary source. Internal evaluations are not listed.

1. DeepSWE. Huang, Lee, Tng, and Ge (Datacurve), 2026. https://deepswe.datacurve.ai/ · https://arxiv.org/abs/2607.07946
2. APEX-SWE (ApexSWE). Kottamasu et al. (Mercor), 2026. https://www.mercor.com/apex/apex-swe-leaderboard/ · https://arxiv.org/abs/2601.08806
3. SWE-Bench Pro. Deng et al. (Scale AI), 2025. https://labs.scale.com/leaderboard/swe_bench_pro_public · https://arxiv.org/abs/2509.16941
4. SWE-Bench Multilingual. Khandpur, Lieret, Jimenez, Press, and Yang (SWE-bench), 2025. https://www.swebench.com/multilingual.html
5. SWE-Marathon. Desai et al. (Abundant AI), 2026. https://www.swe-marathon.org/ · https://arxiv.org/abs/2606.07682
6. ProgramBench. Yang, Lieret, et al., 2026. https://programbench.com/ · https://arxiv.org/abs/2605.03546
7. Terminal-Bench 2.1. Merrill et al., 2026. https://www.tbench.ai/ · https://arxiv.org/abs/2601.11868
8. SWE-Atlas-QnA. Artificial Analysis, 2026. https://artificialanalysis.ai/agents/coding-agents?coding-agents-performance-chart=swe-atlas-qna
9. GDPval-AA (AA GDPval). Artificial Analysis. https://artificialanalysis.ai/evaluations/gdpval-aa
10. τ-bench. Yao, Shinn, Razavi, and Narasimhan (Sierra), 2024. https://taubench.com/ · https://arxiv.org/abs/2406.12045
11. RelBench. Robinson et al. (Stanford), 2024. https://relbench.stanford.edu · https://arxiv.org/abs/2407.20060
12. CyberGym. Wang et al., 2025. https://www.cybergym.io/ · https://arxiv.org/abs/2506.02548
13. VCT. Götting, Medeiros, et al., 2025. https://arxiv.org/abs/2504.16137
14. WMDP. Li et al., 2024. https://wmdp.ai · https://arxiv.org/abs/2403.03218
15. LAB-Bench. Laurent et al., 2024. https://github.com/Future-House/LAB-Bench · https://arxiv.org/abs/2407.10362
16. BixBench. Mitchener, Laurent, et al., 2025. https://github.com/Future-House/BixBench · https://arxiv.org/abs/2503.00096
17. FORTRESS. Knight, Deshpande, et al. (Scale AI), 2025. https://labs.scale.com/leaderboard/fortress · https://arxiv.org/abs/2506.14922
18. MASK. Ren et al., 2025. https://www.mask-benchmark.ai/ · https://arxiv.org/abs/2503.03750