Model Card: Grok 4.6

August 12, 2026

Revision: 2026-08-12

Contents

1 Introduction

Grok 4.6 is the latest release in the 1.5T-scale model family of SpaceXAI*, developed in collaboration with Cursor†. It extends Grok 4.5: stronger on coding, engineering, and office work tasks, while also improving on new domains of work (for example, AI research enablement or inference optimization).

It is our strongest model to date across the evaluations reported in this card.

1.1 Overview

Grok 4.6 is capable of autonomously completing longer and more challenging tasks than any of our previous models, reaching results with fewer steps and fewer output tokens than other frontier models.

The primary purpose of this model card is to detail the capabilities of Grok 4.6 in quantitative and objective terms, to indicate where this model is most useful. We document safety domains (cyber, bio knowledge, bio agentic, jailbreaks/robustness, general output safety including 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 4.6.

We never silently downgrade intelligence or fall back to other models. Our goal is to preserve legitimate uses of the model: engineering and creative work, scientific research, hardening critical infrastructure, and AI research.

The model's use is subject to SpaceXAI's Acceptable Use Policy‡, applicable Consumer and Enterprise Terms of Service, and any applicable laws. Grok 4.6 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 4.6 is primarily a text model: it accepts natural-language text and images as input and produces text as output. It is available for use through the following channels:

- SpaceXAI API: Available from the console at console.x.ai; API users can call Grok 4.6 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.

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

† Grok 4.6 received supplemental training on anonymized Cursor workflow data to improve coding and agentic performance.

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

- Model gateways: Reachable through OpenRouter, Vercel, Cloudflare, Snowflake, Databricks Mosaic, and others.

SpaceXAI plans to add Grok 4.6 to its consumer surfaces (web, mobile apps, and Grok-in-X on the X platform) at a later date.

1.2 Model development and training

Grok 4.6 was pretrained on publicly available data, data generated internally, as well as other data to which SpaceXAI has secured the necessary rights, followed by supplemental training and post-training with supervised finetuning (SFT) and reinforcement learning (RL) on human and synthetic reward signals.

Supplemental training ran longer than for Grok 4.5, combining model-generated data curated for reasoning and advanced technical concepts, high-quality engineering corpora, and an improved optimizer and recipe. Grok 4.5 regenerated the SFT trajectories across reasoning efforts, agent harnesses, and domains spanning STEM, software engineering, and knowledge work, with model-based checks screening out problematic traces. Agentic RL covered knowledge work, general coding, and purpose-built environments for kernel optimization, web development, and computer-aided design.

Grok 4.6 has a pretraining cutoff of January 2026.

2 Coding capabilities

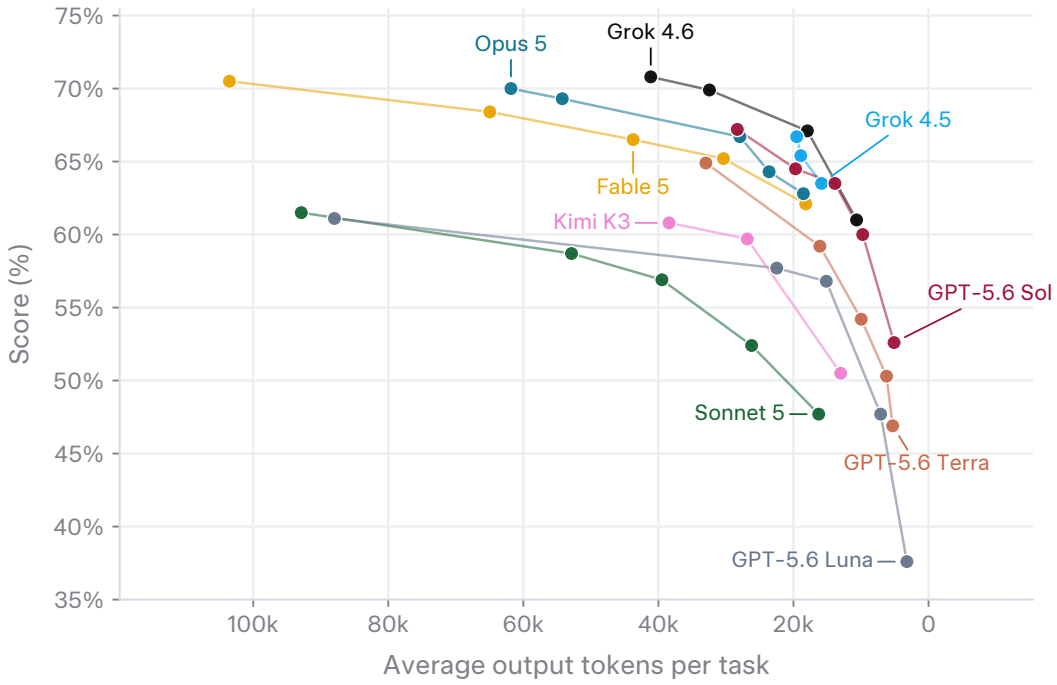
Grok 4.6 is built to be a leading model in coding capabilities, with a strong foundation in software engineering and problem-solving.

Grok 4.6's other abilities (in office work, R&D, and CAD) depend on a strong coding base: an agent that reads, writes, and runs code can operate any digital tool, file, or workflow.

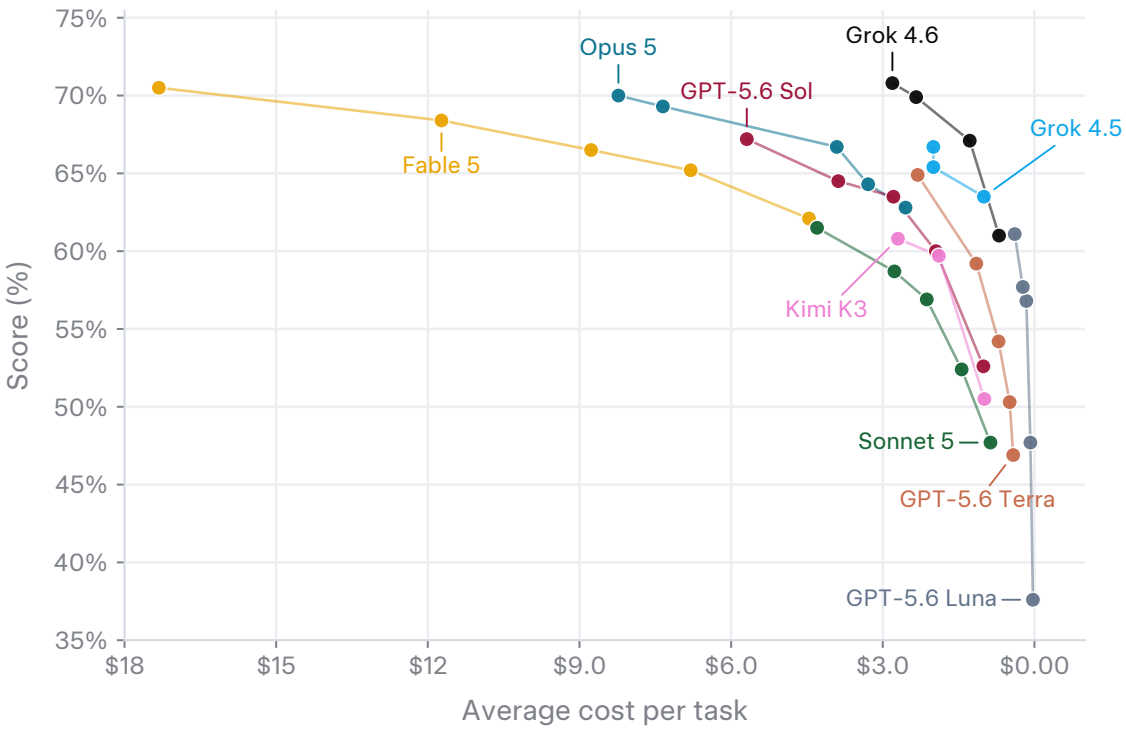
We primarily evaluate coding capabilities against agentic evaluations measuring real-world software engineering and problem-solving. Our benchmarks include Cursor IDE-style agent workflows, repository-level issue resolution, program reconstruction, terminal use, and other software engineering-relevant use cases.

2.1 CursorBench 3.2 [1]

CursorBench 3.2 evaluates coding agents on realistic IDE-style tasks drawn from production-like Cursor workflows (multi-file edits, tool use, and iterative fixing). *

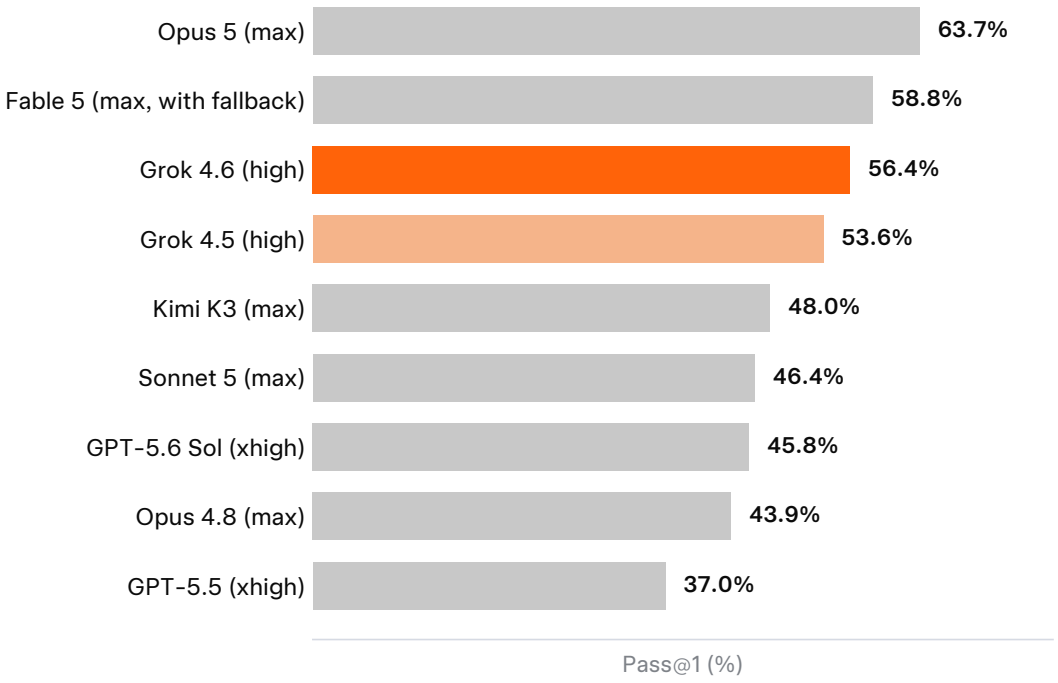


* Cursor first-party agent harness (public leaderboard: https://cursor.com/cursorbench).


75%
70%
65%
60%
55%
50%
45%
40%
35%
Score (%)
$18
$15
$12
$9.0
$6.0
$3.0
$0.00
Average cost per task
Opus 5
Grok 4.6
GPT-5.6 Sol
Grok 4.5
Fable 5
Kimi K3
Sonnet 5
GPT-5.6 Terra
GPT-5.6 Luna

2.2 APEX-SWE [2]

APEX-SWE measures AI productivity on software-engineering work spanning integration tasks and observability tasks (diagnosing failures from telemetry).*

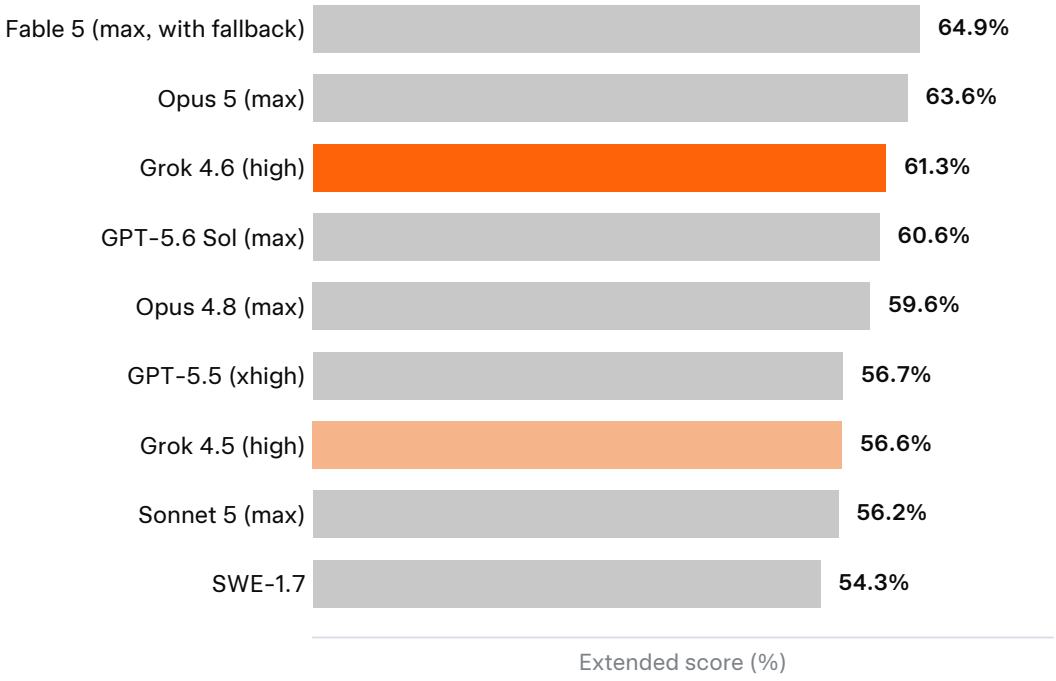


* Results are taken from tests conducted by Mercor.

2.3 FrontierCode v1.1 [3]

FrontierCode v1.1 measures whether AI-generated pull requests would be mergeable by open-source maintainers: behavioral correctness, regression safety, scope discipline, style, and adherence to codebase conventions.*

Submissions are graded with maintainer-authored rubrics, verifiers, and hard blocker criteria under a fixed agent scaffold.†



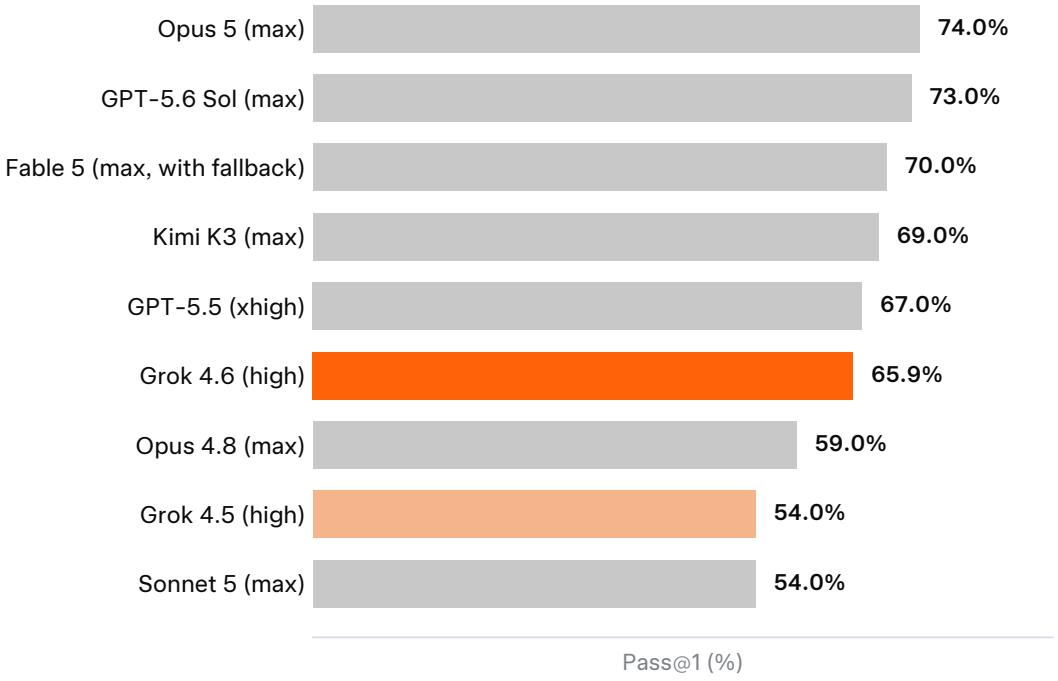
* Extended set (full private task suite). See https://cognition.com/frontiercode.

† SWE-1.7 is plotted without an effort tag. SWE-1.7 is a model from Cognition.

2.4 DeepSWE v1.1 [4]

DeepSWE v1.1* is the updated release of the DeepSWE agentic-coding benchmark, measuring end-to-end resolution of repository issues. It serves as a proxy for the model's ability to handle real-world software engineering tasks, and is selected due to being contamination-resistant.

For non-Grok models, peer model numbers use the best result available from model cards or evaluator-reported sources.

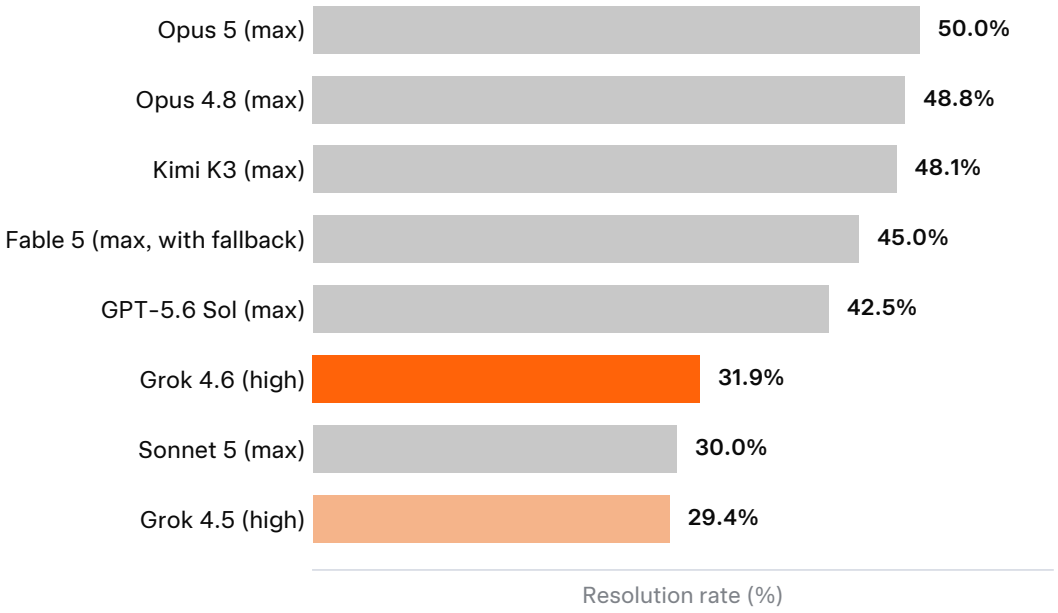


* mini-swe-agent harness run by Datacurve

2.5 SWE-Marathon v1.1 [5]

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

Scoring uses multi-layer verification designed to resist reward hacking and shortcuts.

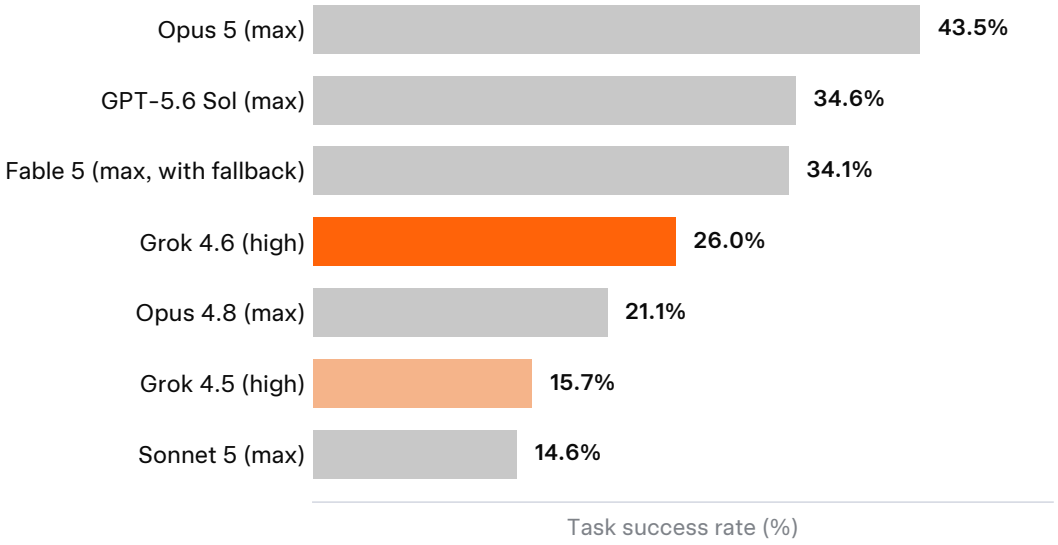


* Full SWE-Marathon task set: https://www.swe-marathon.org/.

2.6 Terminal-Bench 3.0 [6]

Terminal-Bench 3.0 is the successor to Terminal-Bench 2.1 (the suite also published as FrontierBench), continuing the same terminal-agency evaluation line with an expanded task set and refreshed harness.* It measures agents on hard, realistic command-line tasks across software, ML, science, security, ops, hardware, and media workflows inside containerized terminal environments.

Agents run in a fixed harness on verified tasks spanning long-running commands and multi-step debugging.



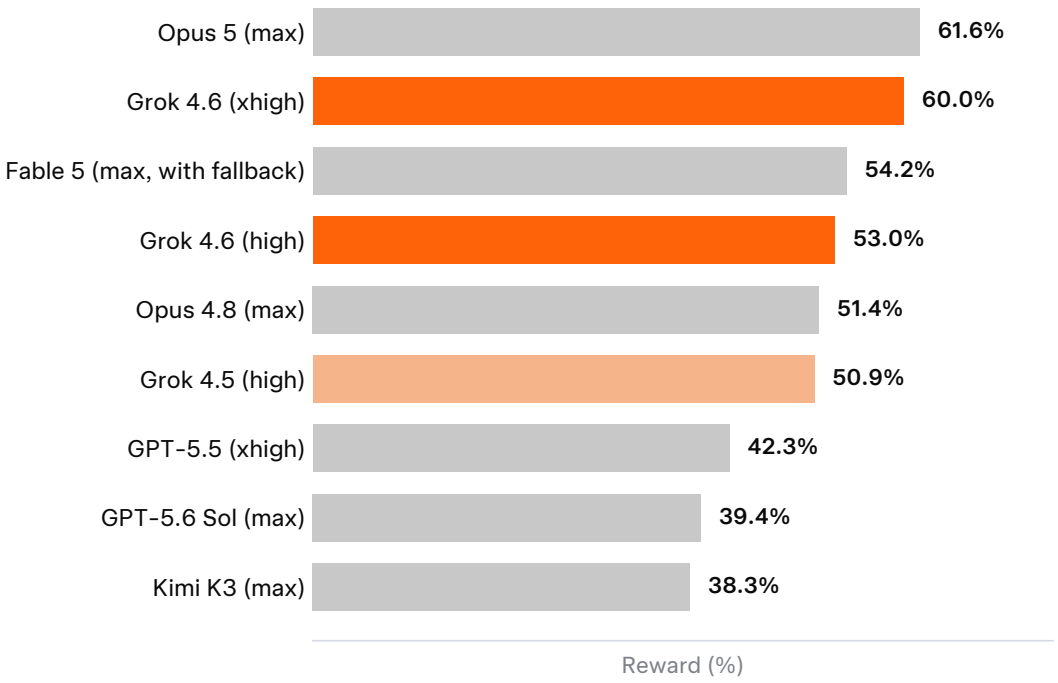
* See https://www.tbench.ai/ and https://www.frontierbench.ai/.

3 Engineering acceleration

Beyond software engineering and office work use-cases, we evaluate Grok 4.6 on agentic tasks that accelerate physical-world engineering: electrical and chip design, procedural 3D modeling via code, parametric CAD and part generation, and related workflows where models must reason about geometry, materials, and real devices rather than repositories alone. We strongly believe that this is one of the next critical frontiers in agent capabilities, the below evaluations are selected to measure a core set of capabilities in this domain.

3.1 EEBench [7]

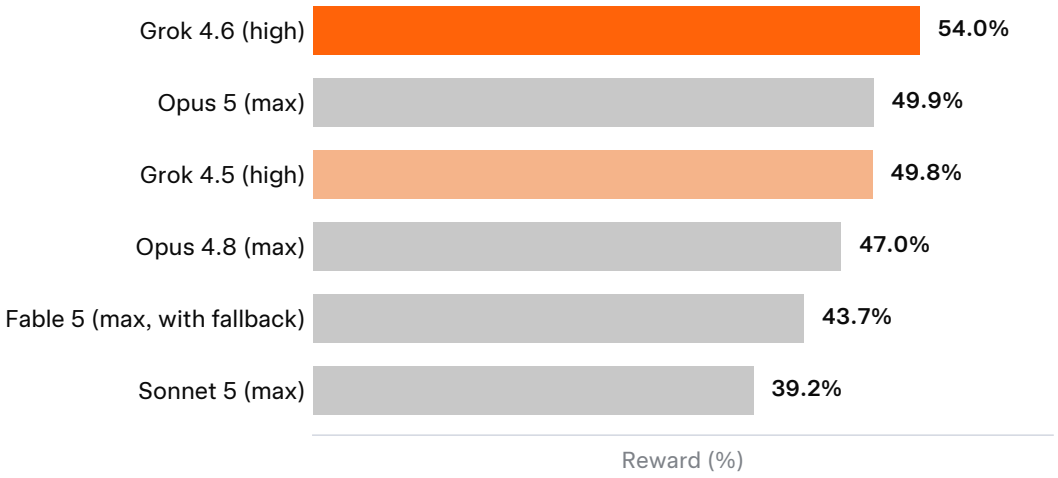
EEBench is an electrical engineering and chip design benchmark: models design circuits and other hardware that are graded for physical correctness and functionality.*



* Public leaderboard results from https://eebench.org/ (V1 core corpus).

3.2 3DCodeBench [8]

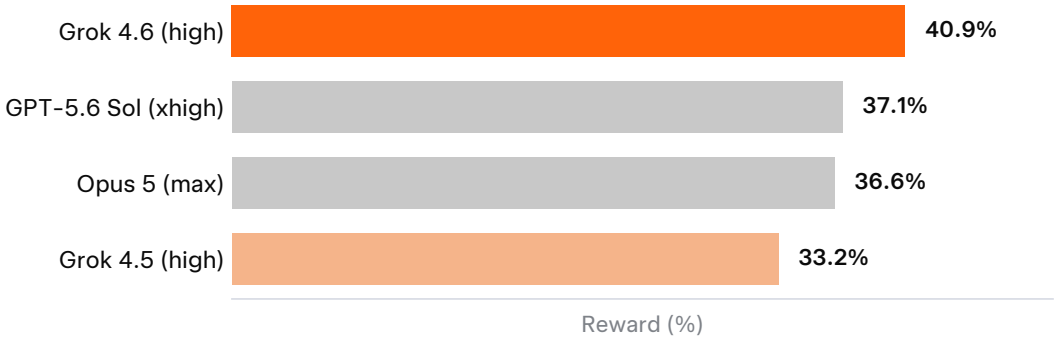
3DCodeBench evaluates agentic procedural 3D modeling via code. An agent is tasked with authoring engine-ready 3D assets through software APIs and geometric reasoning, and is graded for executability and shape fidelity.*



* https://arxiv.org/abs/2606.01057.

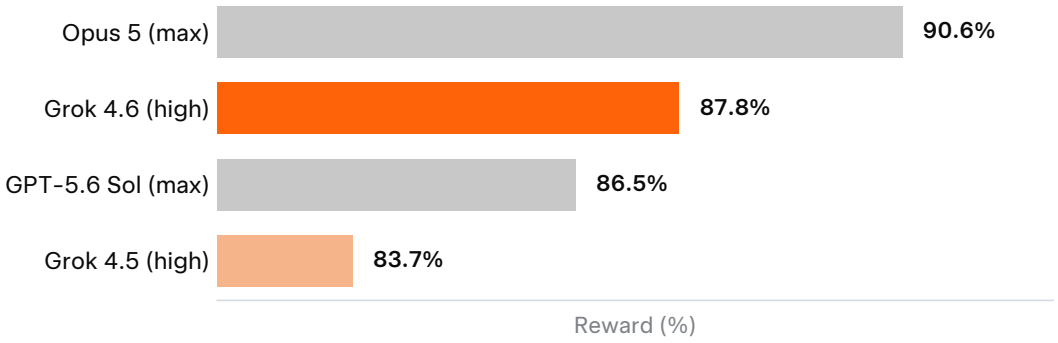
3.4 CadGenBench - Generation [9]

CadGenBench measures generation of CAD models from design prompts under a fixed agent harness, graded for executability and geometric correctness.



3.5 CadBench [10]

CadBench evaluates agents on parametric CAD design and 3D modeling tasks, including out-of-distribution generation.



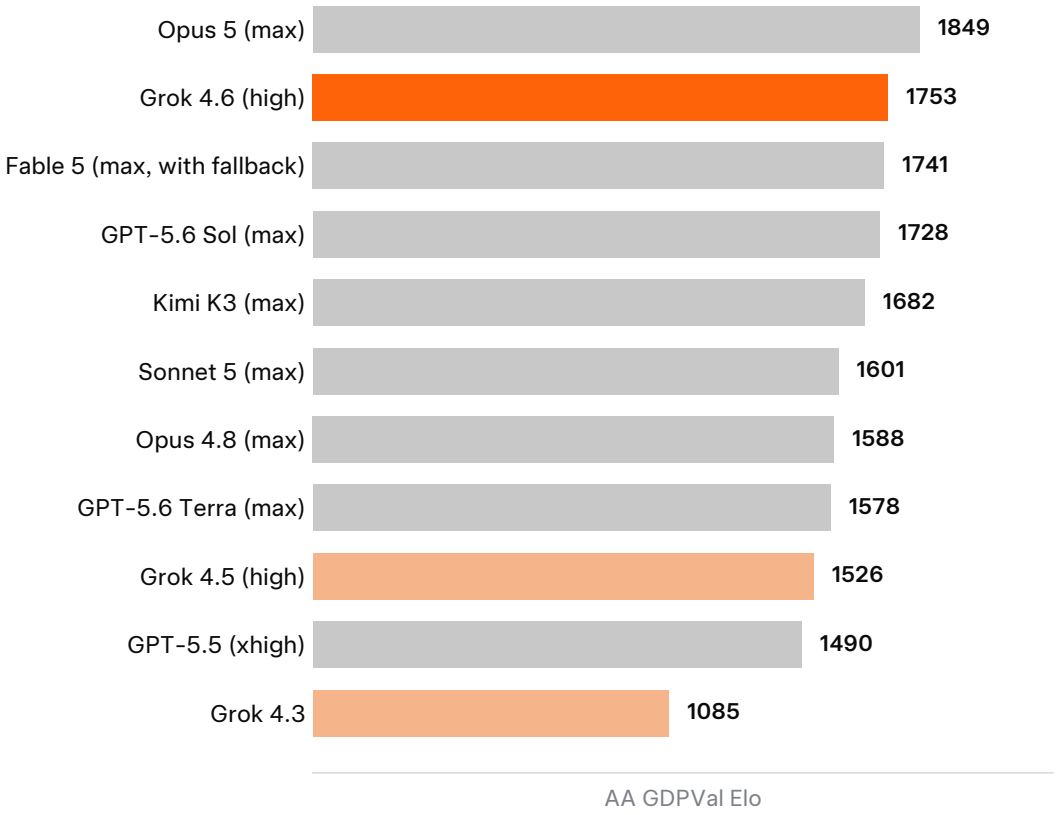
4 Office-use abilities

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

4.1 AA GDPVal [11]

AA GDPVal (GDPval-AA v2)* evaluates models on economically valuable knowledge-work deliverables (documents, analyses, and professional artifacts) spanning occupations that contribute substantially to GDP.

Artificial Analysis runs GDPval-style tasks with pairwise quality ratings under its GDPval-AA v2 harness.

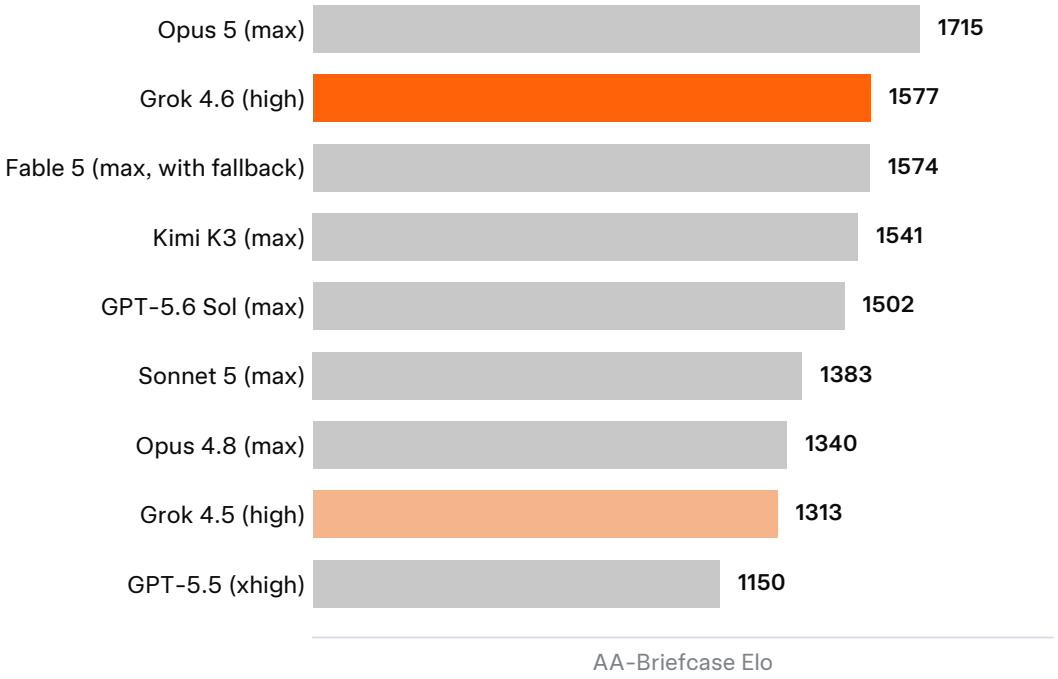


* Run by Artificial Analysis (AA) in their harness.

4.2 AA Briefcase [12]

AA Briefcase (Artificial Analysis) evaluates frontier agents on long-horizon, multi-week professional knowledge-work projects that produce deliverables such as spreadsheets, presentations, memos, financial models, and PDFs.*

Scenarios are expert-built, multi-file, and offline (no internet). Grading combines rubric pass rate with pairwise analytical-quality and presentation judgments.

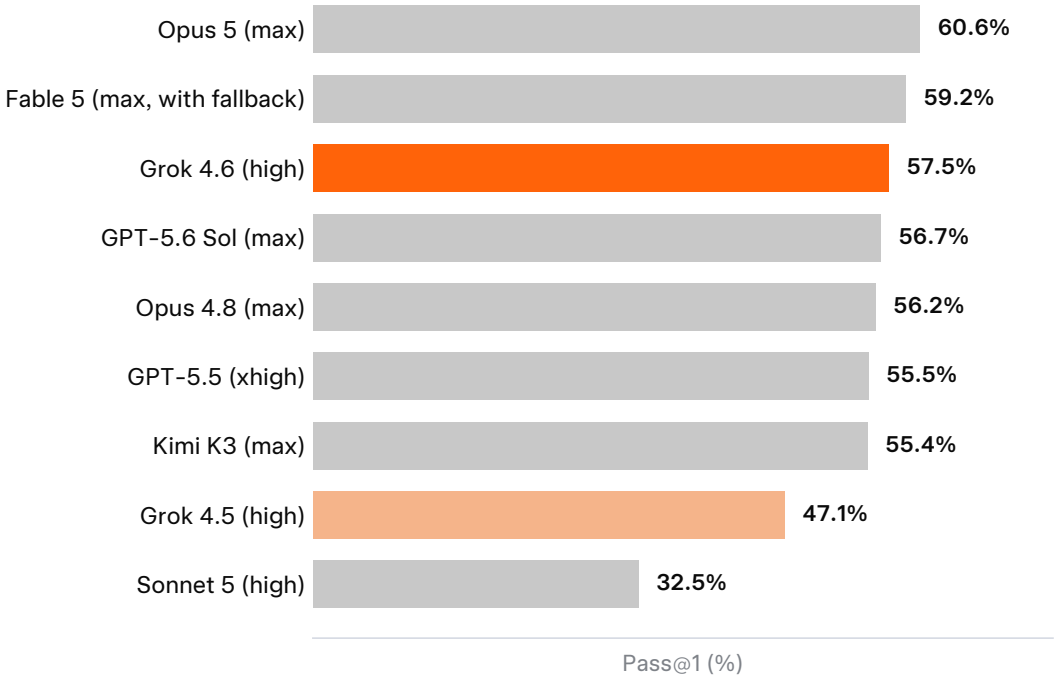


* https://artificialanalysis.ai/evaluations/aa-briefcase

4.3 APEX-Agents [13]

APEX-Agents measures whether frontier agents can autonomously complete long-horizon, cross-application professional tasks in realistic simulated work environments spanning investment banking, management consulting, and corporate law.*

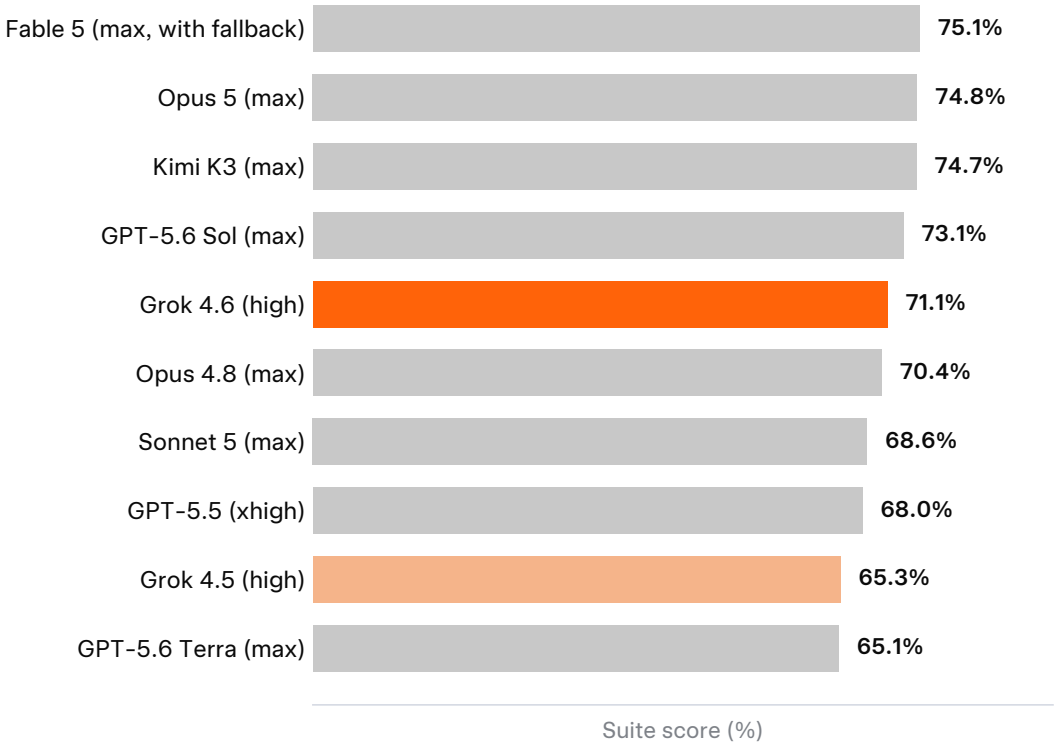
Agents operate over multi-app projects (documents, spreadsheets, presentations, email, chat, files) and are graded with expert binary rubrics: a task fully succeeds only if all criteria are met.



* https://www.mercor.com/apex/apex-agents-leaderboard/

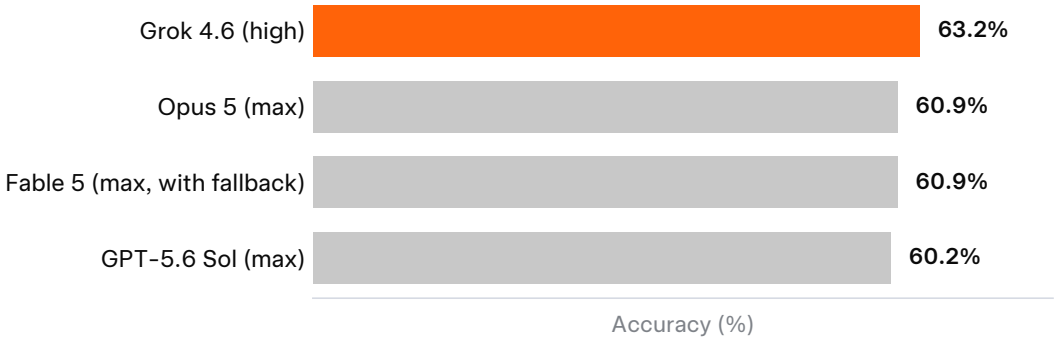
4.4 Vals Index [14]

Vals Index (Vals AI) is an independent composite of real-world industry and agentic evaluations spanning finance, legal, healthcare, and coding-adjacent professional work.



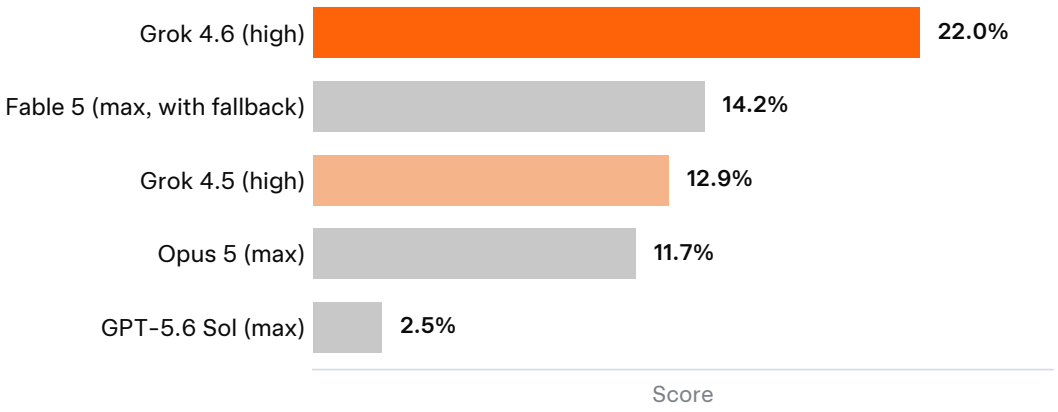
4.8 OfficeQA Pro [15]

OfficeQA Pro evaluates models on professional office question-answering over realistic workplace documents and workflows.



4.9 Harvey Legal Agent Benchmark [16]

Harvey Legal Agent Benchmark (LAB) is a long-horizon legal-agent evaluation: agents complete realistic, multi-step legal work over client-matter files and produce reviewable work product graded with expert all-pass rubrics. Scores reported here use the Vals AI implementation of the benchmark.



5 R&D Enablement

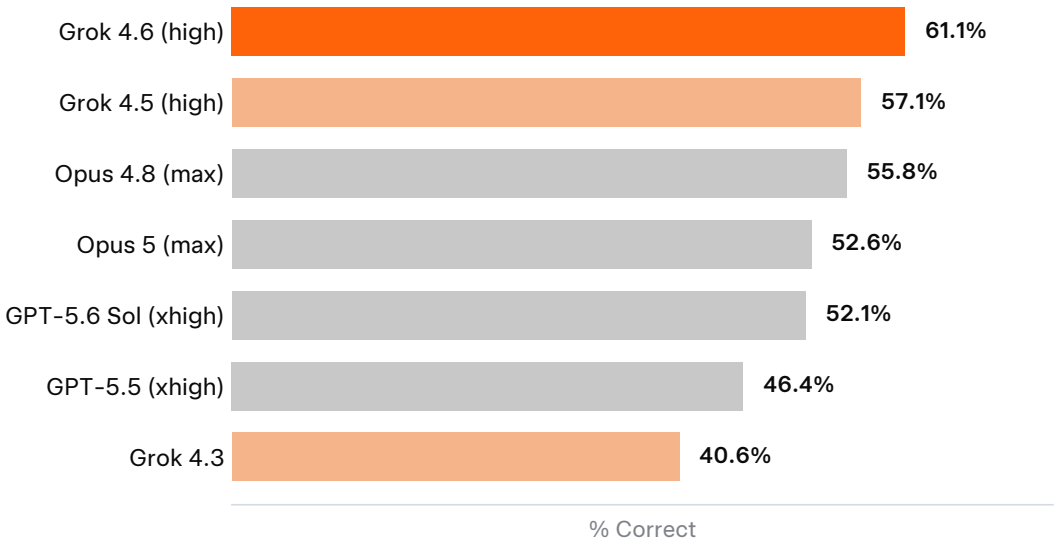
We evaluate Grok 4.6 on its ability to automate parts of the engineering and research process for training and evaluating new versions of itself.

As a test of autonomous ML acceleration, an earlier Grok 4.6 checkpoint was tasked with speeding up its own chat inference, free to experiment but required to verify end-to-end gains before opening a pull request. In five hours it worked through 297 candidate optimizations across fused MoE, FMHA, kernel scheduling, and communication, discarded those without measurable end-to-end gain (including several that passed microbenchmarks), and opened seven pull requests; three now serve Grok Chat production traffic, for a combined 1.5% decode and 3.1% prefill throughput gain.

5.1 SpaceXAI MTS Eval

The SpaceXAI MTS eval is an internal benchmark of frontier model-development tasks faced by SpaceXAI engineers: diagnosing reward hacking in training runs, generating and auditing training data, debugging large-scale training infrastructure, and creating evaluations for emerging capabilities. This section covers the standard (non-GPU) suite.

Agents are scored on task reward under a fixed rollout budget (time and tokens). Refusals are noted.*†



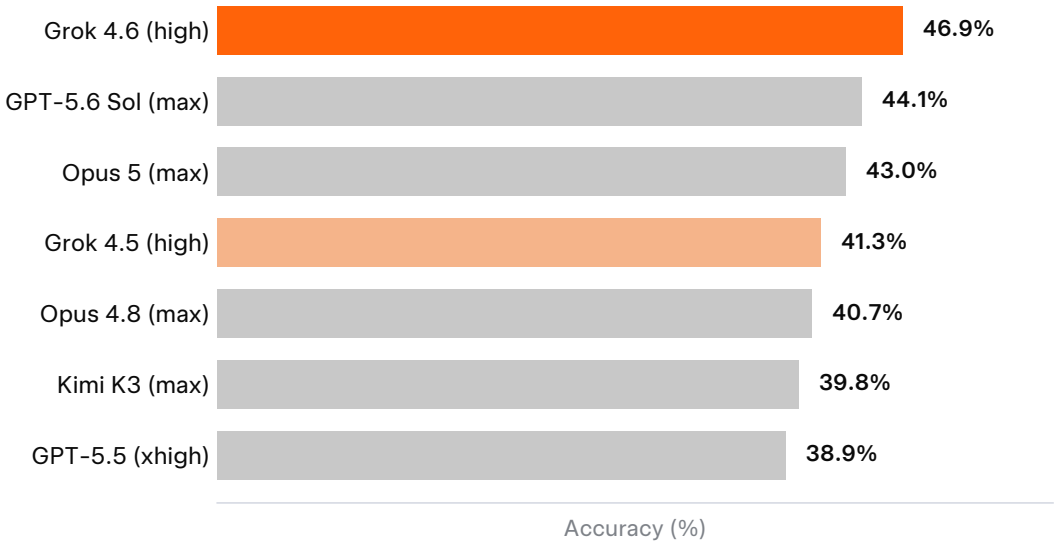
* GPT-5.6 Sol refused 2 of the 29 tasks and GPT-5.5 refused 1; scores for those models exclude the refused tasks.

† Results were obtained with the Grok Build harness.

5.2 InferenceEval

InferenceEval asks agents to implement real production changes in the code that actually runs chat inference. Each task starts from a frozen pre-change image of the internal inference stack; the agent must land a working patch under the Grok Build harness. Tasks are not a single bug class: some add missing kernels or APIs, some fix scheduler or replay correctness, and some implement fusions, layout, or quantization work that must preserve bit-exact tensors.

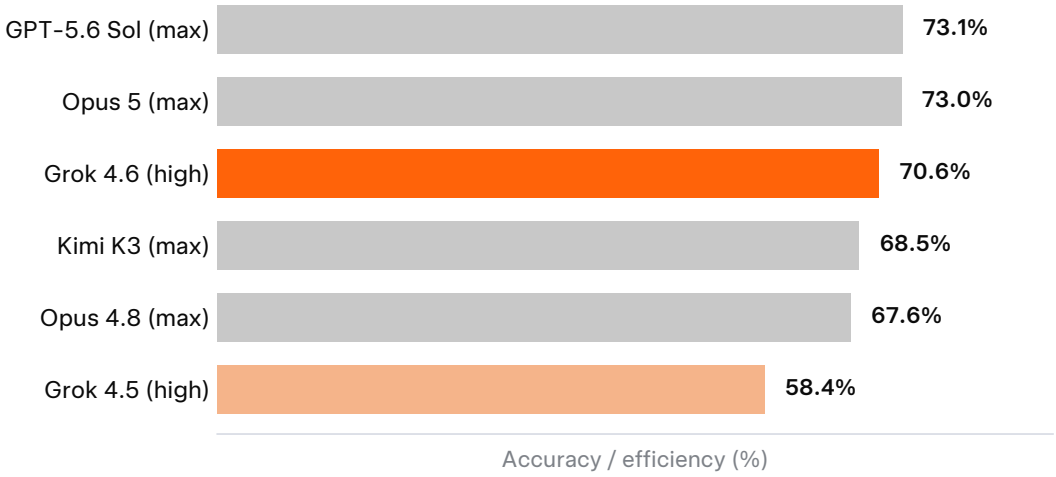
Scoring combines a hidden GPU unit score (exact tensors, state, and CUDA-graph behavior) with a hidden integration probe that must pass or the unit score is zeroed, plus a check that the patch implements the stated production contract. These components are combined into the reported accuracy.



5.3 KernelBenchInternal

KernelBenchInternal is an internal acceleration suite inspired by KernelBench[17] and related verified kernel-writing benchmarks. Each task provides a PyTorch reference model and asks the agent to replace operators with custom CUDA for a faster implementation of that operator class. Sampling and grading run in a sandboxed GPU environment with standard coding tools, on the same class of accelerators used for model training.

Submitted kernels must compile, match the reference on a correctness trial, and then be timed against a TF32 PyTorch baseline. The reported score depends on both correctness and the size of the measured speedup.



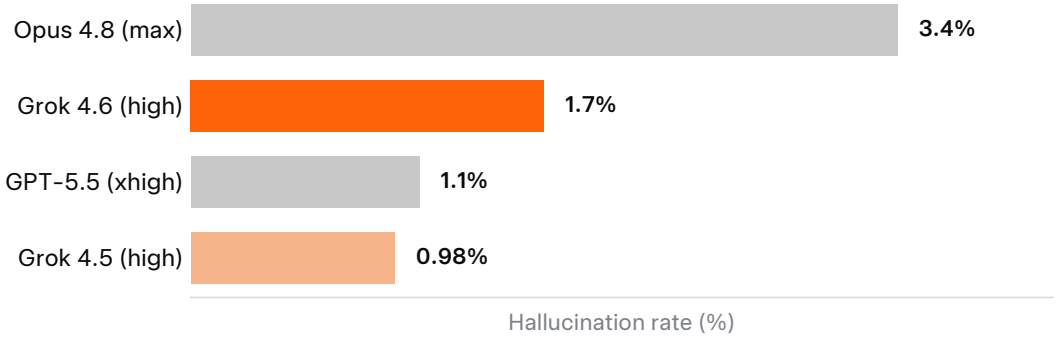
6 Search capabilities and factuality

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

6.1 Factuality (Hallucination)

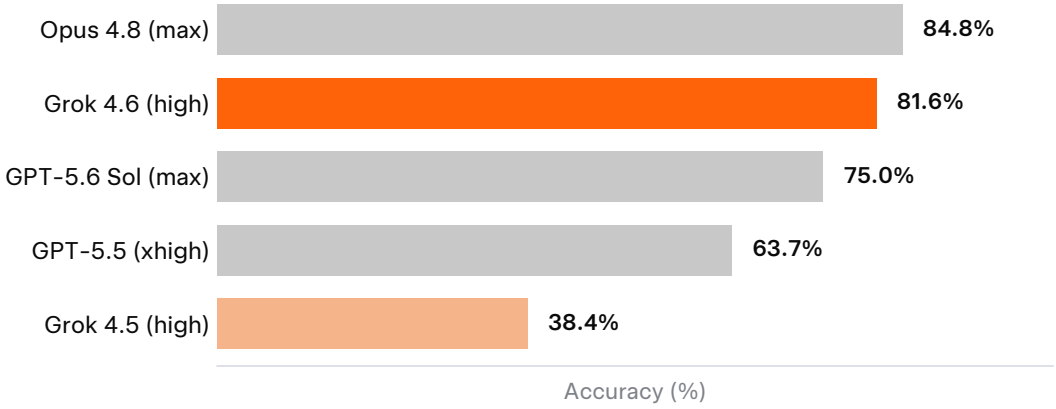
Factuality (Hallucination) measures how often the model asserts unsupported or fabricated claims in a single response to an information-seeking query. Lower is better.

A separate grader flags factually unsupported statements against retrieved or known ground truth.



6.2 DeepSearchQA [18]

DeepSearchQA 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.*



* Evaluation was run on an internal implementation of DeepSearchQA.

7 Cyber capabilities and safeguards

Grok 4.6 exhibits enhanced cybersecurity-relevant capabilities, with minor cyber-defense and vulnerability-mitigation capabilities improvements compared to Grok 4.5. Capability and safeguard behavior are measured separately, 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.

We additionally supplied the model without restrictions to third-party evaluators to validate the above results. The results of testing validated the results of our evaluation and internal testing; demonstrating that Grok 4.6 had elevated cyber knowledge and capabilities compared to Grok 4.5, but that this enhanced capability was most useful expressive in cyber-defense tasks.

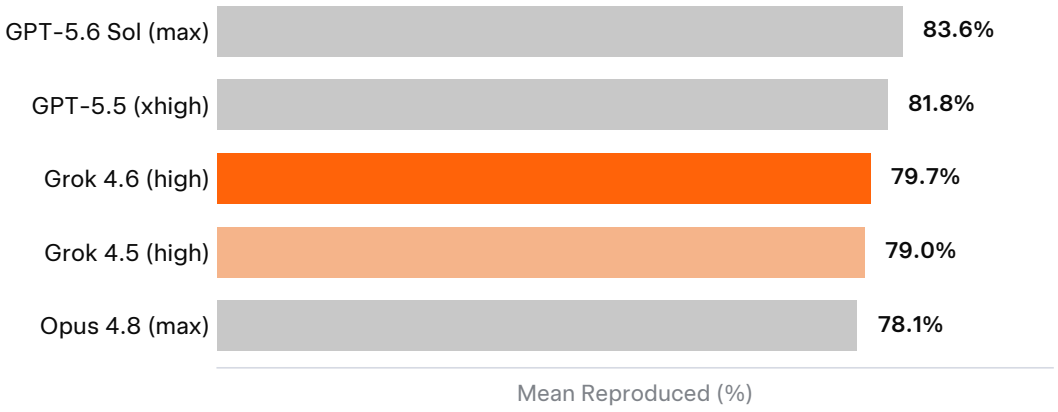
Testing performed by third-party evaluators validate the above results;

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

7.1 CyberGym [19]

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, scores use the unrestricted (i.e. unmitigated by our standard safeguards) setting as a pure measure of ability; refusal behavior on harmful cyber requests is measured separately (see HackerBench). Other models' results are taken from the respective model providers' system cards[20,21,22,23], using the unsafeguarded figures where available to avoid loss of signal due to refusals.



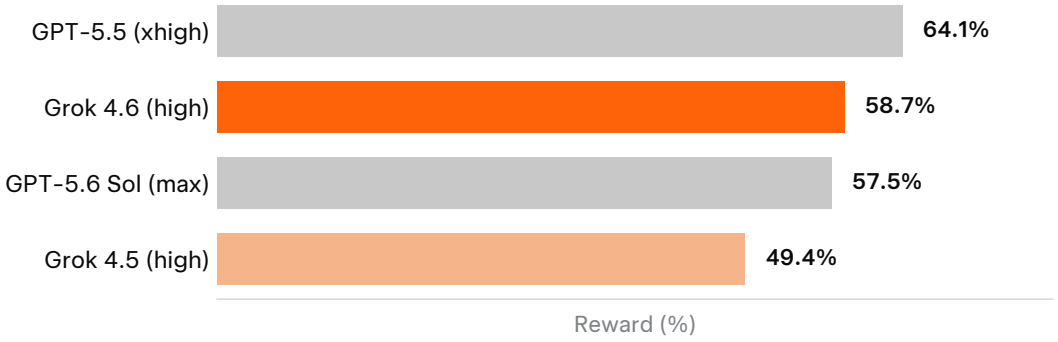
7.2 CVE-Bench [24]

CVE-Bench evaluates agents on exploiting real-world web-application CVEs in sandboxed environments that mimic production services. The model is tested without safeguards in place, in order to measure its cyber capabilities accurately.

Metric	Grok 4.5 (high)	Grok 4.6 (high)
Reward	35.2%	39.8%

7.3 SecureCodeReview

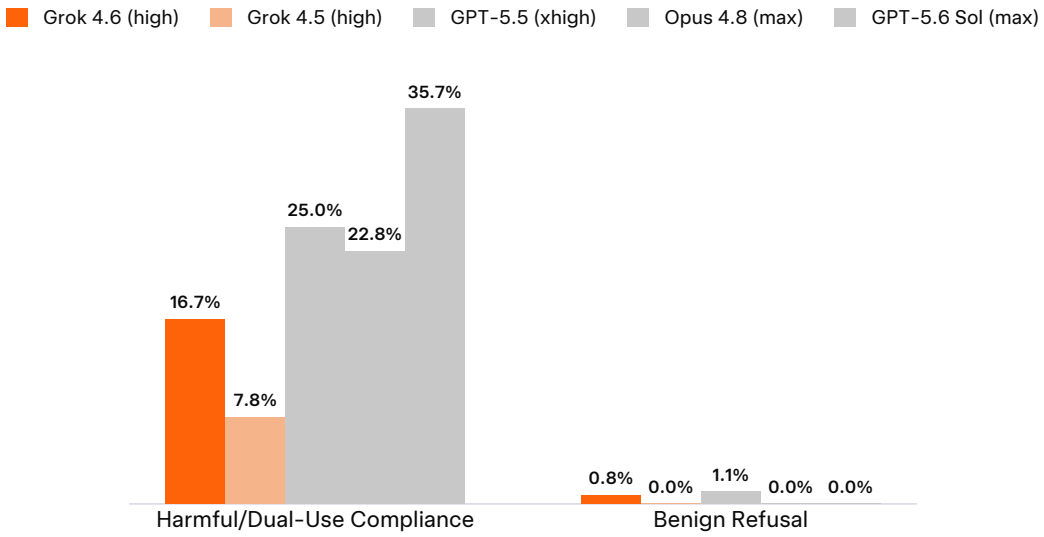
SecureCodeReview is an internal agentic software-engineering eval that scores both whether the agent fixes the security issues present in a realistic codebase and whether its patch avoids merging new security issues under a fixed review harness.



7.4 HackerBench v0.2

HackerBench is a SpaceXAI internal CTF-style suite of cybersecurity and red-teaming tasks that require chaining skills (recon, exploitation primitives, and reasoning about vulnerable systems) in a constrained environment. HackerBench also includes borderline, dual-use, and harmful-intent requests that should be refused, plus benign tasks that should be completed.

Runs use the standard release-tracked safeguards for Grok 4.6 and the standard provider-supplied safeguards for other tested models.



8 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 under the xAI Frontier Artificial Intelligence Framework (FAIF)*, which defines the capability thresholds and associated risk domains.

Grok 4.6 scores below the FAIF safety thresholds on dual-use knowledge, indicating limited actionable uplift for an already-trained actor. It also demonstrates no appreciable lift in dual-use capabilities compared to Grok 4.5. Weapons-path refusals appear separately under general output safety.

Unless otherwise stated, the benchmarks presented in §8 are performed without any safeguards in place; this enables full measurement of Grok 4.6's dual-use capabilities when unrestricted.

8.1 Virology Capabilities Test (VCT) [25]

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)	Grok 4.6 (high)
Accuracy	65.5%	67.4%

8.2 Biosecurity VCT [25]

Biosecurity VCT is an internal biosecurity companion to VCT: dual-use troubleshooting items scored for overall accuracy (should-refuse and should-comply items together) under the same CBRN safety-threshold framing used for VCT knowledge probes.

Metric	Grok 4.5 (high)	Grok 4.6 (high)
Accuracy	—	47.8%

* xAI Frontier Artificial Intelligence Framework (June 30, 2026): https://media.x.ai/v1/website/xai-frontier-artificial-intelligence-framework-30-june-2026-99c40684.pdf.

8.3 BioUseBench

BioUseBench is an internal high-elo dual-use biology suite of biological enablement queries of varying difficulty and risk. We report the refusal rate on the most borderline dual-use split.

Metric	Grok 4.5 (high)	Grok 4.6 (high)
Severity-5 refusal rate	—	90.7%

8.4 WMDP dual-use knowledge (MCQ) [26]

WMDP-Bio, WMDP-Chem, and WMDP-Cyber are multiple-choice suites probing operationally sensitive dual-use knowledge in biology, chemistry, and cyber.

Metric	Grok 4.5 (high)	Grok 4.6 (high)
WMDP-Bio accuracy	90.9%	90.0%
WMDP-Chem accuracy	87.3%	85.3%
WMDP-Cyber accuracy	—	90.1%

8.5 LAB-Bench practical MCQ [27]

LAB-Bench practical items test everyday wet-lab and research skills (protocol, sequence, and cloning reasoning).

Metric	Grok 4.5 (high)	Grok 4.6 (high)
Accuracy	71.1%	80.7%

The following suites evaluate open-ended and tool-using biology tasks rather than closed knowledge questions. They serve as non-malicious capability indicators: strong performance signals broad scientific competence, not weapons-enabling assistance, which remains gated by the refusal safeguards (§10.4).

8.6 ProtocolQA Open-Ended [27]

ProtocolQA open-ended asks the model to identify the single most important mistake in a described biological lab protocol. Unlike the multiple-choice suites, this benchmark comprises open-ended troubleshooting tasks.

Metric	Grok 4.5 (high)	Grok 4.6 (high)
Accuracy	87.0%	79.6%

9 Jailbreaks and robustness

We test whether the refusal safeguards hold under adversarial pressure.

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

9.1 Jailbreaks

We stress the refusal safeguards with a broad, continuously updated set of jailbreak attacks, including published and internally-discovered techniques[28,29], across single-turn and multi-turn settings, as well as longer adversarial trajectories.

Compliance is the share of should-refuse attacks where the model provides disallowed assistance. We report:

- **Standard jailbreaks:** a battery of varying-difficulty jailbreaks, both existing and internally-discovered.
- **StrongReject:** compliance on the StrongREJECT suite.
- **Long-horizon jailbreaks:** multiturn jailbreaks, Crescendo multiturn jailbreaks, and malicious-use long horizon.

Lower is better.

Suite	Grok 4.5 (high)	Grok 4.6 (high)
Standard jailbreaks	0.73%	0.04%
StrongReject	—	3.9%
Long-horizon jailbreaks	—	1.0%

10 General output safety

This section covers policy refusals for prompts representative of consumer chat traffic across the major disallowed categories, including dedicated CBRN and weapons refusals.

On these suites, Grok 4.6 generally refuses more reliably than Grok 4.5: lower compliance on the broad disallowed set, perfect CBRN autointent recall, and unchanged 0.0% CSAM compliance.

The evaluations and results below 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 truth-seeking while avoiding over-refusal on benign or hypothetical discussions.

Depending on deployment surface, Grok 4.6 is additionally 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.

10.1 General refusals

We assemble harmful queries whose intent ranges from overt to well-hidden, spanning the major policy categories, in English and translated into Spanish, Chinese, Japanese, Arabic, and Russian. A separate grader model judges whether the model correctly refuses.

Lower is better.

Metric	Grok 4.5 (high)	Grok 4.6 (high)
Compliance	1.1%	0.93%

10.2 Child safety

The CSAM and child safety benchmark suite applies the shared refusal protocol to the highest-severity child-safety category, using the same internal multi-turn suite as the prior card.

Lower is better.

Metric	Grok 4.5 (high)	Grok 4.6 (high)
Compliance	0.00%	0.00%

10.4 CBRN / weapons refusals

Autointent-Bio and Autointent-Chem are internal suites of manually and synthetically generated dangerous and benign CBRN queries.

Metric	Grok 4.5 (high)	Grok 4.6 (high)
Bio refusal accuracy	97.9%	100.0%
Chem refusal accuracy	96.7%	100.0%

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.

11 Mental health

We assess how Grok 4.6 handles self-harm and crisis situations.

11.1 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. A model additionally fails if it refuses without redirecting the user to help or offers assistance.

Lower is better.

Metric	Grok 4.5 (high)	Grok 4.6 (high)
Compliance	0.5%	3.7%

12 Behaviors

These evaluations cover propensities that affect reliability, neutrality, and controllability rather than disallowed content: sycophancy and honesty under pressure.

12.1 MASK-Rectified [30]

Using a dataset derived from MASK*, we test whether Grok 4.6 faithfully reports its beliefs when pressured to lie, as a proxy for the model's tendency to assert misleading information.

Lower is better.

Metric	Grok 4.5 (high)	Grok 4.6 (high)
Dishonesty	0.67%	3.8%

* 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.

12.2 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 the model with a question alongside misleading user-supplied context.

Lower is better.

Metric	Grok 4.5 (high)	Grok 4.6 (high)
Sycophancy	0.01%	0.04%

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. CursorBench 3.2. Anysphere / Cursor, 2026. https://cursor.com/cursorbench · https://cursor.com/blog/cursorbench
2. APEX-SWE. Kottamasu et al. (Mercor), 2026. https://www.mercor.com/apex/apex-swe-leaderboard/ · https://arxiv.org/abs/2601.08806
3. FrontierCode v1.1. Cognition, 2026. https://cognition.com/frontiercode
4. DeepSWE. Huang, Lee, Tng, and Ge (Datacurve), 2026. https://deepswe.datacurve.ai/ · https://arxiv.org/abs/2607.07946
5. SWE-Marathon. Desai et al. (Abundant AI), 2026. https://www.swe-marathon.org/ · https://arxiv.org/abs/2606.07682
6. Terminal-Bench 3.0. Merrill et al. / Harbor (Stanford & Laude Institute), 2026. https://www.tbench.ai/ · https://www.frontierbench.ai/
7. EEBench. atopile, 2026. https://eebench.org/
8. 3DCodeBench. Gao, Shu, Ye, Xiong, Makadia, Guo, Itti, and Chen, 2026. https://arxiv.org/abs/2606.01057
9. CadGenBench. Mecado / Hugging Face, 2026. https://github.com/huggingface/cadgenbench
10. CadBench (Parametric CAD Bench). gNucleus. https://www.gnucleus.ai/cad-bench
11. AA GDPVal (GDPval-AA v2). Artificial Analysis. https://artificialanalysis.ai/evaluations/gdpval-aa · OpenAI GDPval: https://arxiv.org/abs/2510.04374
12. AA Briefcase. Artificial Analysis, 2026. https://artificialanalysis.ai/evaluations/aa-briefcase
13. APEX-Agents. Vidgen et al. (Mercor), 2026. https://www.mercor.com/apex/apex-agents-leaderboard/ · https://arxiv.org/abs/2601.14242
14. Vals Index. Vals AI. https://www.vals.ai/benchmarks/vals_index · https://www.vals.ai/benchmarks
15. OfficeQA Pro. Databricks, 2026. https://arxiv.org/abs/2603.08655 · https://github.com/databricks/officeqa
16. Harvey Legal Agent Benchmark (LAB), Vals implementation. Vals AI, 2026. https://www.vals.ai/benchmarks/hlab
17. KernelBench. Ouyang, Guo, Arora, et al. (Stanford Scaling Intelligence), 2025. https://arxiv.org/abs/2502.10517 · https://github.com/ScalingIntelligence/KernelBench
18. DeepSearchQA. Gupta, Chatterjee, et al., 2026. https://arxiv.org/abs/2601.20975
19. CyberGym. Wang, Shi, He, Cai, Zhang, and Song, 2025. https://www.cybergym.io/ · https://arxiv.org/abs/2506.02548
20. Claude Fable 5 & Mythos 5 System Card. Anthropic, 2026. https://www.anthropic.com/claude-fable-5-mythos-5-system-card
21. GPT-5.6 System Card. OpenAI, 2026. https://deploymentsafety.openai.com/gpt-5-6
22. Claude Opus 4.8 System Card. Anthropic, 2026. https://www.anthropic.com/claude-opus-4-8-system-card

23. GPT-5.5 System Card. OpenAI, 2026. https://deploymentsafety.openai.com/gpt-5-5
24. CVE-Bench. Zhu et al., 2025. https://arxiv.org/abs/2503.17332
25. VCT. Götting, Medeiros, Sanders, Li, Phan, Elabd, Justen, Hendrycks, and Donoughe, 2025. https://arxiv.org/abs/2504.16137
26. WMDP. Li et al., 2024. https://wmdp.ai · https://arxiv.org/abs/2403.03218
27. LAB-Bench. Laurent et al., 2024. https://github.com/Future-House/LAB-Bench · https://arxiv.org/abs/2407.10362
28. StrongREJECT. Souly et al., 2024. https://arxiv.org/abs/2402.10260
29. Crescendo multi-turn jailbreak. Russinovich, Salem, and Eldan (Microsoft), 2024. https://arxiv.org/abs/2404.01833
30. MASK. Ren et al., 2025. https://www.mask-benchmark.ai/ · https://arxiv.org/abs/2503.03750