Model Card: Grok 4.7

September 21, 2026

Revision: 2026-09-21

Contents

1 Introduction

Grok 4.7 is SpaceXAI's* latest model. It extends Grok 4.6, demonstrating greater capability and autonomy on coding†, engineering, and office work tasks.

It is our most capable model to date.

1.1 Overview

Grok 4.7 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 4.7 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 4.7.

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[1], applicable Consumer and Enterprise Terms of Service, and any applicable laws. Grok 4.7 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.

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

† Grok 4.7 received supplemental training on anonymized Cursor workflow data to improve coding and agentic performance.

Grok 4.7 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 4.7 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 Grok by SpaceXAI add-ins for Microsoft Word, PowerPoint, and Excel.
- Model gateways: Reachable through OpenRouter, Vercel, Cloudflare, Snowflake, Databricks Mosaic, and others.

SpaceXAI plans to add Grok 4.7 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 4.7 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 fine-tuning (SFT) and reinforcement learning (RL) on human and synthetic reward signals.

Supplemental training for Grok 4.7 ran longer than for Grok 4.6, combining model-generated data curated for reasoning and advanced technical concepts, high-quality engineering corpora, and an improved optimizer and recipe. Grok 4.6 models were used to generate 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 4.7 has a pretraining data cutoff of June 2026, and uses data generated as late as August 2026 in its supplemental training.

2 Coding capabilities

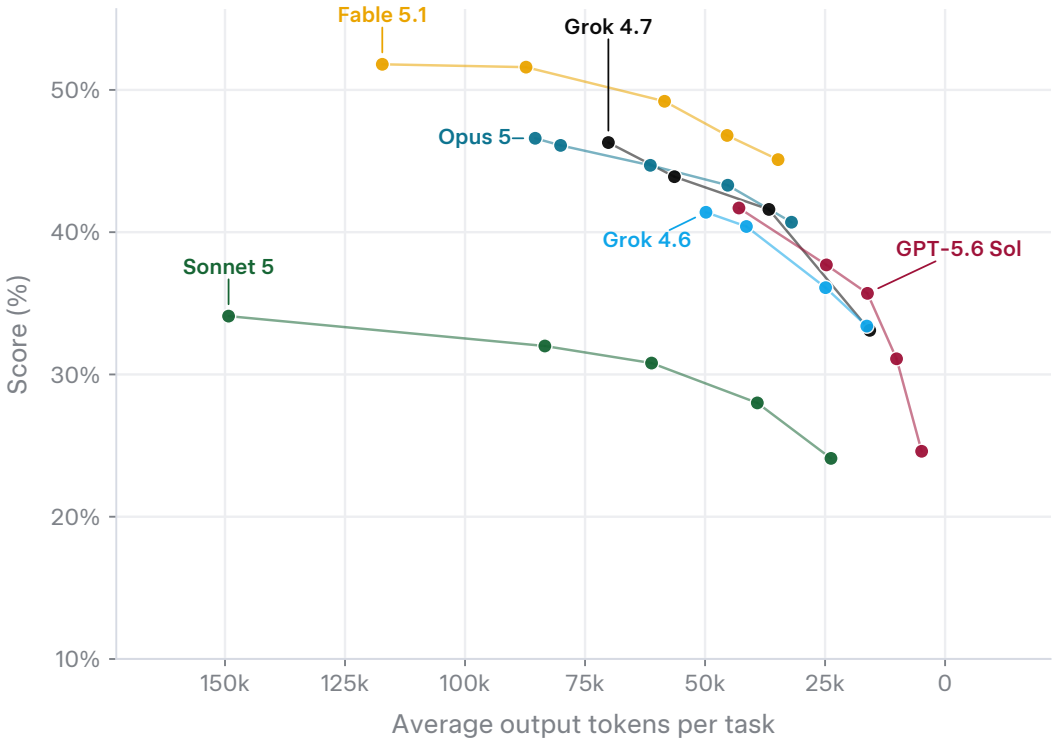
Grok 4.7 is built for strong coding performance across wide-ranging domains of software engineering, problem-solving, and software-engineering-adjacent tasks.

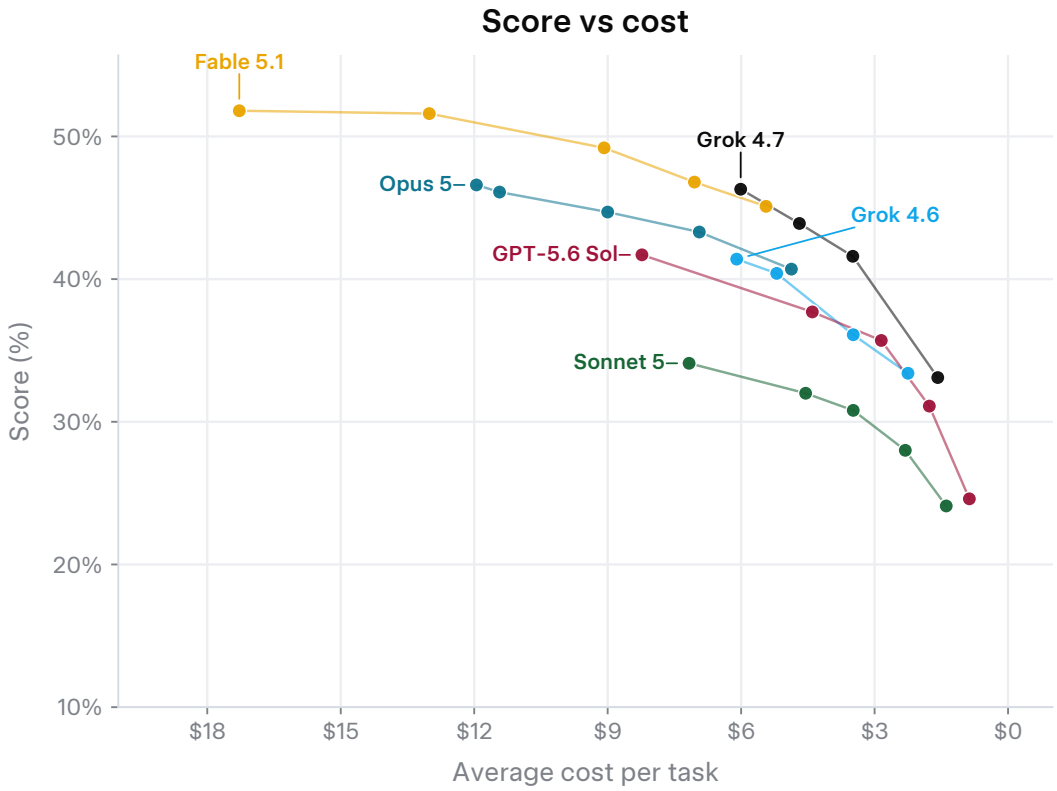
Grok 4.7's other abilities (in office work, engineering, and CAD) depend on well-developed baseline coding capabilities: an agent that reads, writes, and runs code is a prerequisite for expressive interaction with any digital tool, file, or workflow.

We primarily evaluate coding on agentic benchmarks of real-world software engineering, problem-solving, and software-engineering-adjacent tasks, such as SRE and observability. These benchmarks cover Cursor IDE-style agent workflows, repository-level issue resolution, terminal use, and other software-engineering-relevant use cases.

2.1 CursorBench 4.0 [2]

CursorBench 4.0 evaluates coding agents on long-horizon coding tasks from real Cursor sessions. Version 4.0 adds longer-horizon tasks, so its scores are not comparable with CursorBench 3.2. Grok 4.7 improves on long-horizon coding tasks by thoroughly considering edge cases, working persistently on difficult problems, and self-verifying more effectively. Grok 4.7 scores 46.3% at xhigh and 43.9% at high.

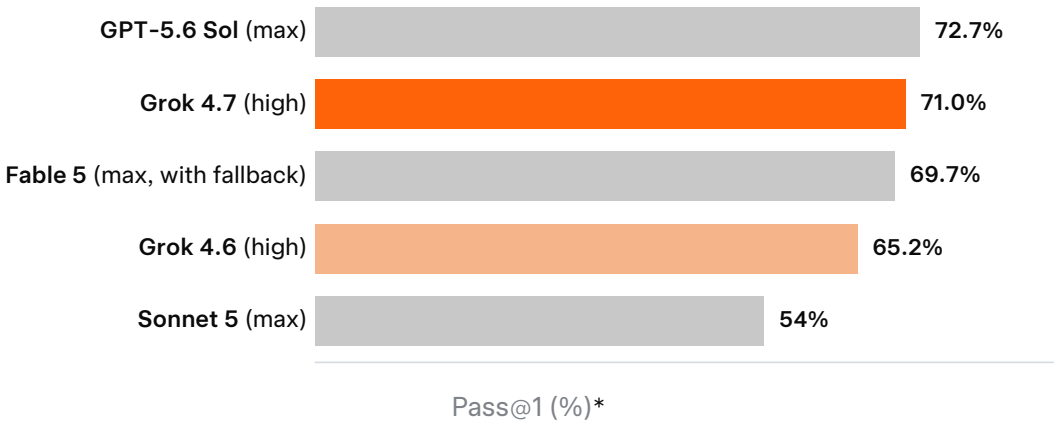



Score vs cost
Fable 5.1
Grok 4.7
Opus 5–
Grok 4.6
GPT-5.6 Sol–
Sonnet 5–
50%
40%
30%
20%
10%
Score (%)
$18
$15
$12
$9
$6
$3
$0
Average cost per task

2.2 DeepSWE v1.1 [3]

DeepSWE v1.1 evaluates coding agents on 113 original, long-horizon software engineering tasks across 91 active repositories and five languages. Tasks are written from scratch rather than mined from existing commits or pull requests, keeping their reference solutions out of the public record. Hand-written verifiers test the requested behavior and accept alternative correct implementations.

Every model runs through the mini-SWE-agent harness. Each committed patch is extracted, applied to a pristine verifier container, and graded independently of the agent's runtime environment. Grok 4.7 scores 71.0% at high.



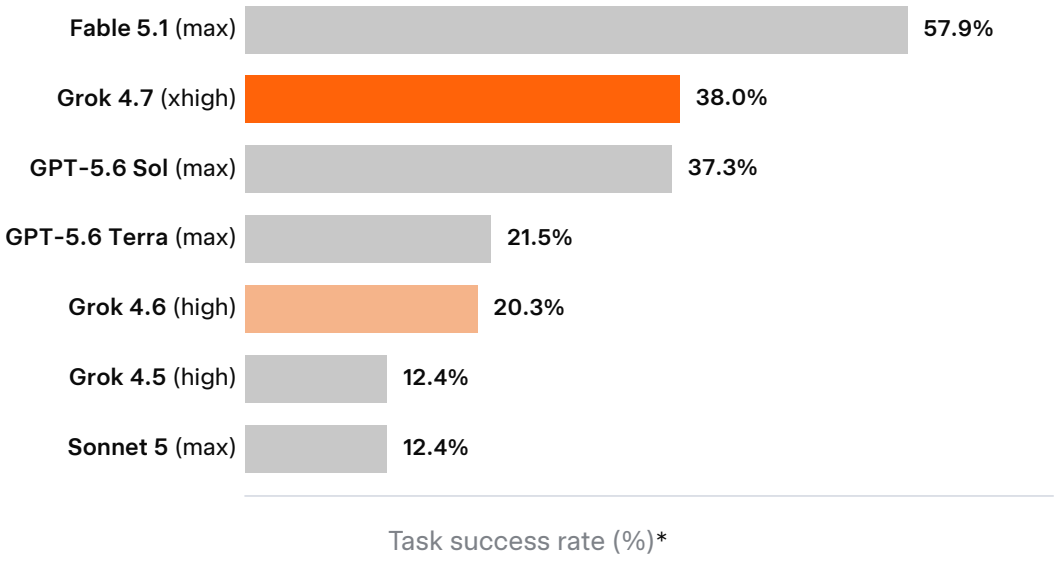
* Results reported are taken from evaluations conducted by Datacurve.

2.3 Terminal-Bench 4.0 [5]

Terminal-Bench 4.0 consists of 66 difficult technical workflows in real terminal environments, spanning software, machine learning, science, operations, security, hardware, and media. Each task has a programmatic verifier. Version 4.0 recalibrates CPU, memory, and time budgets, gives agents up to eight hours, fixes unstable tasks, and removes tasks that were saturated.

Grok 4.7 scores 38.0% at xhigh, evaluated with the Grok Build harness. Grok 4.6 scores 20.3% at high.

The resolution rate measures end-to-end terminal execution, including planning, tool use, recovery, and verification. The recalibrated resources reduce infrastructure-driven failures, but absolute scores remain sensitive to the agent harness.

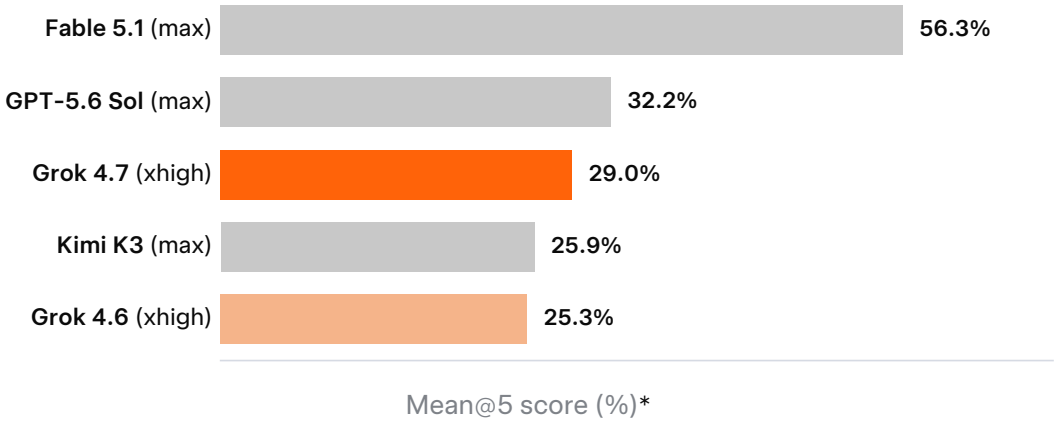


* Results reported are taken from evaluations conducted by Harbor.

2.4 FrontierSWE V2 [6]

FrontierSWE V2 evaluates coding agents on 34 ultra-long-horizon challenges spanning systems implementation, performance engineering, scientific computing, visual reasoning, and AI research. Agents receive up to 20 hours per task, and task-specific verifiers award scores from zero to one so meaningful partial progress is visible when complete solutions remain rare. The public protocol runs five trials per task and reports mean@5 across the task set.

The headline score is an average partial-credit reward, not task resolution or FrontierSWE V1 Dominance. Public V2 results use Proximal's Proximus harness at maximum effort. Grok 4.7 scores 29.0% at xhigh, evaluated by Proximal Labs; Grok 4.6 scores 25.3% at xhigh.

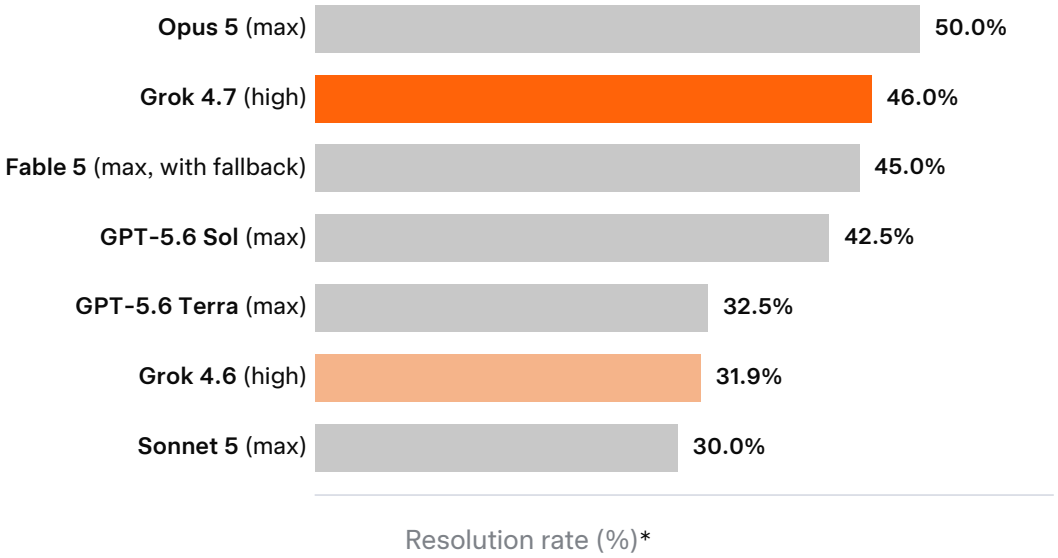


* Results are taken from evaluations conducted by Proximal Labs.

2.5 SWE-Marathon v1.1 [7]

SWE-Marathon v1.1 evaluates autonomous completion of 20 realistic, multi-hour software engineering tasks spanning library reproductions, product clones, machine-learning work, and performance optimization. Task-specific verifiers combine hidden tests, behavioral parity checks, performance gates, integrity checks, and computer-use review where interface quality matters.

The benchmark runs eight trials per task, and a trial counts as resolved only when every verifier passes. Because models run in their native agent harnesses, scores reflect the complete model-agent system. Grok 4.7 scores 46.0% at high.



* Results reported are taken from evaluations conducted by Abundant AI.

3 Knowledge-work capabilities

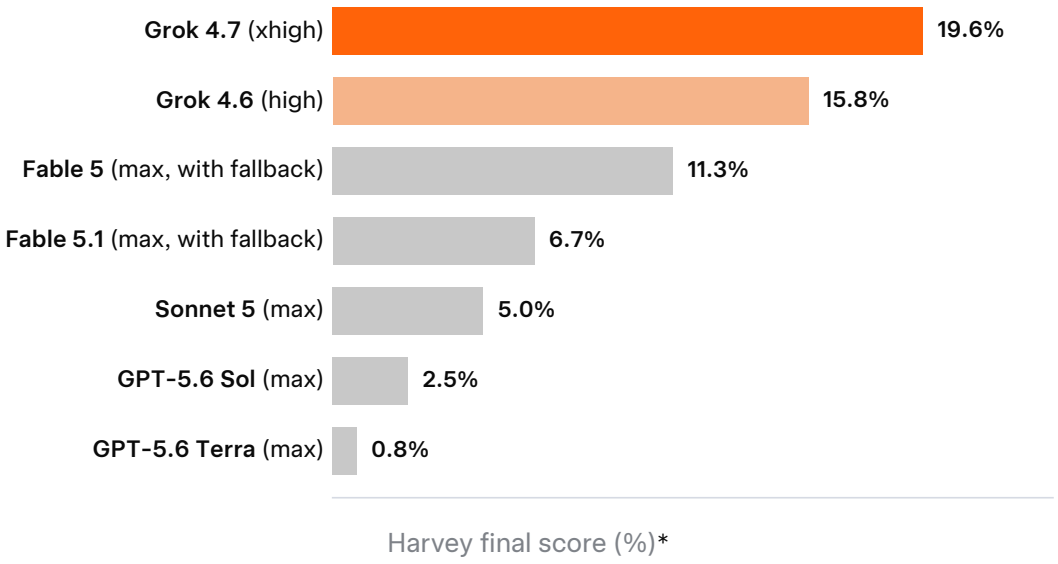
We evaluate performance on professional knowledge-work and structured agent tasks relevant to office and productivity use cases, in order to quantify Grok 4.7's utility in assisting, accelerating, and automating knowledge work. We select these benchmarks for task- and domain-diversity, as well as for alignment with subjective user experiences of performance (that is, a higher score is consistent with a material improvement in the utility of the tested model and the quality of produced artifacts).

On all tested domains, Grok 4.7 demonstrates frontier or near-frontier performance.

3.1 Legal Agent Benchmark [8]

The Legal Agent Benchmark evaluates long-horizon legal work organized around realistic client matters. Agents use six file and terminal tools plus document, presentation, and spreadsheet skills to produce work against task-specific criteria.

We report Vals AI's held-out 120-task run in the Valkyrie harness with internet access disabled. Harvey's final score averages two judges' task pass rates; each judge marks a task as passed only when every criterion passes. Grok 4.7 scores 19.6% at xhigh.



* Results are taken from evaluations conducted by Vals AI.

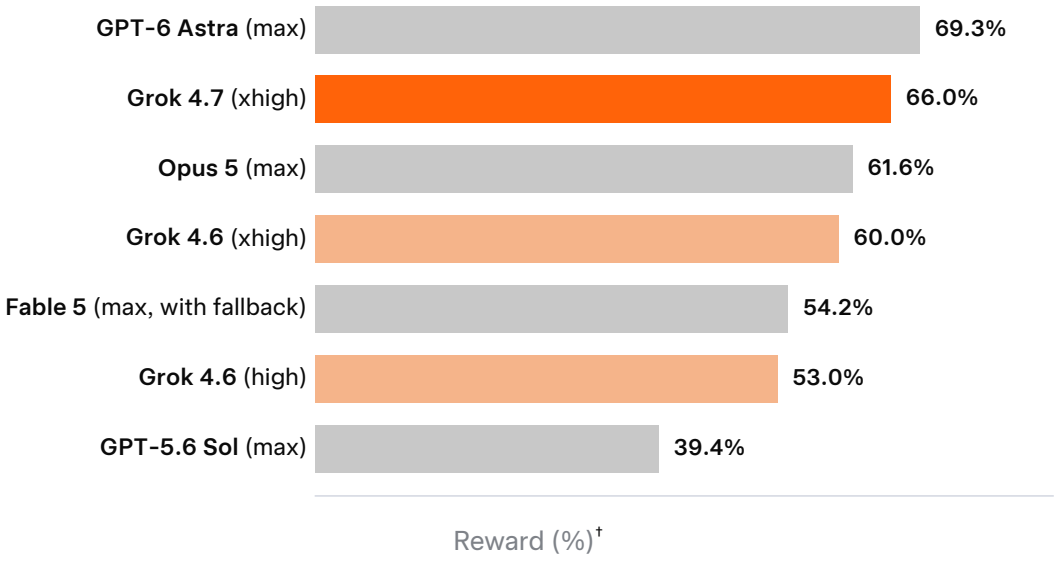
4 Engineering acceleration

Engineering represents one of the core domains where improved agent capabilities directly support the acceleration of technological development: agents that accelerate rocket design, IC layout, and datacenter power-and-cooling optimization compress the timelines of progress across the physical systems that enable further advances in AI capabilities and utility.

We evaluate Grok 4.7 on agentic tasks that accelerate physical-world engineering (i.e. all engineering not purely within the software domain): electrical and chip design, parametric CAD generation, and related workflows where models must reason about geometry, materials, and real devices rather than repositories alone.

4.1 EEBench [10]

EEBench is an electrical engineering and chip design benchmark: models design circuits and other hardware that are graded for physical correctness and functionality.* Grok 4.7 is evaluated with the Grok Build harness; peer models use their respective provider harnesses. Grok 4.7 scores 66.0% at xhigh, compared with 60.0% for Grok 4.6 at xhigh. GPT-6 Astra (max) scores 69.3%. Opus 5 (max) scores 61.6%.

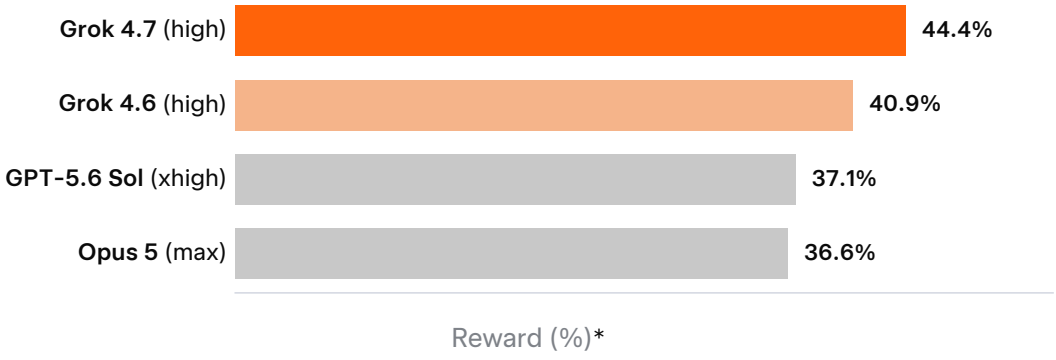


* As part of the V1 core corpus of the benchmark.
† Results are taken from evaluations conducted by Atopile.

4.2 CADGenBench [11]

CADGenBench evaluates agents on CAD model construction from design prompts. We report the generation split of the benchmark: the model must produce CAD geometry graded for executability and geometric correctness.

Grok 4.7 scores 44.4% at high, evaluated with the Grok Build harness; peer models use their respective provider harnesses.



* Results are taken from evaluations conducted by Mecado.

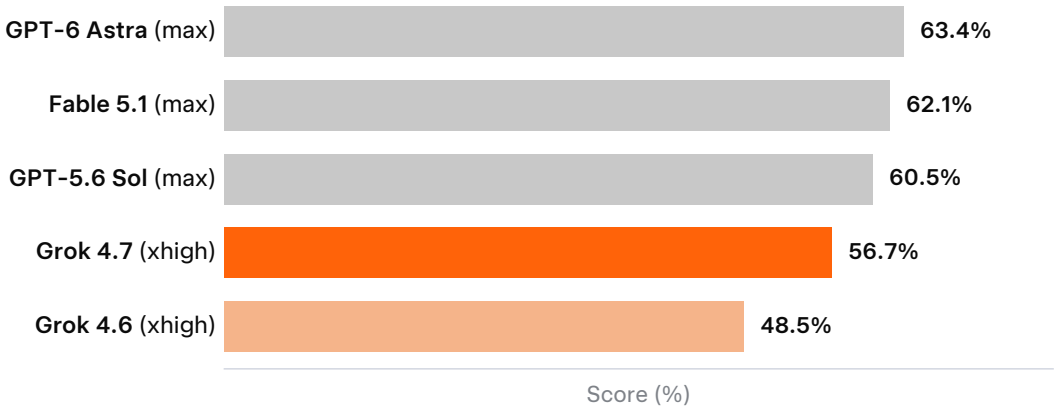
5 Medical and biological capabilities

This section covers clinical communication and agentic biological data analysis: HealthBench Professional for realistic healthcare conversations, and the LatchBio capability benchmark suite.

5.1 HealthBench Professional

HealthBench-Professional is an open-source benchmark for measuring safety, accuracy, and communication in realistic healthcare settings. Grok 4.7 scores 56.7% at xhigh on HealthBench Professional, up from 48.5% for Grok 4.6 at xhigh. It improves on complex clinical reasoning and gives more complete, relevant answers, handling nuanced clinical constraints more accurately.

Grok 4.6 (high) was used as the grader model for Grok 4.7 results; Fable and Astra results are from providers' respective model cards. Tasks that were rejected by safety filters were scored as zero.

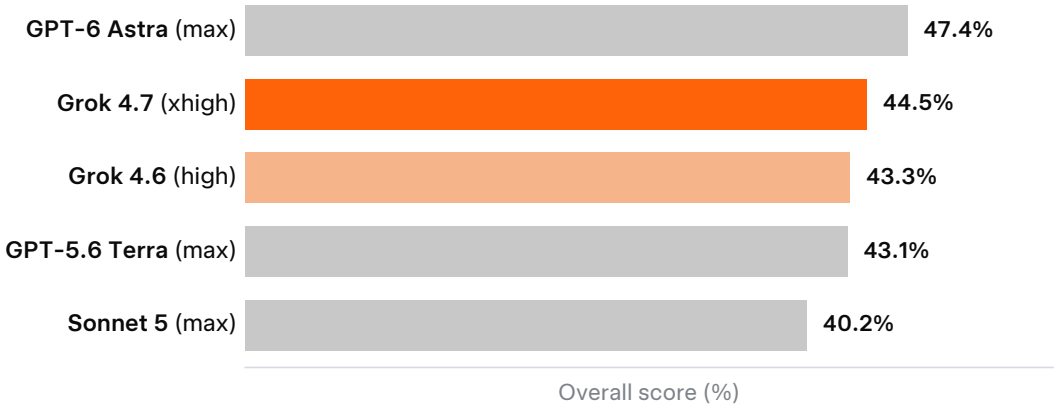


5.2 LatchBio Capabilities v1.0

The following benchmark suite is LatchBio's cross-benchmark evaluation of agentic biological data analysis. Its capability suite spans transcriptomics, epigenomics, variant analysis, therapeutic discovery, preclinical pharmacology, pathogen surveillance, and biological-function inference. Agents inspect real experimental files and return structured conclusions that are evaluated with deterministic, task-specific graders.

The overall score is the equally weighted mean of the observed scores across 11 capability benchmarks. Grok 4.7 scores 44.5% at xhigh.

This is evidence of end-to-end biological analysis rather than recall of biological facts. The LatchBio capability benchmark suite is a live suite, so the model card should freeze the evaluated snapshot and coverage; its Surveillance and Function components also appear in the separately reported BioSecBench results and should be treated as drill-downs rather than independent evidence.



6 Cyber capabilities and safeguards

Models with advanced capability in coding and coding-adjacent (terminal-use) domains also, in the absence of safeguards or controls, demonstrate enhanced abilities in defensive and offensive cybersecurity. In this section, we discuss the results of the evaluation suite we use to measure our models' cyber capabilities, and additionally discuss the effectiveness and calibration of our safeguard stack.

Grok 4.7 shows small cybersecurity-relevant capability gains over Grok 4.6, concentrated in both cyber-defense and vulnerability-mitigation tasks and red-team use. Capability tests are run without the safeguards we use in production, which would otherwise mask the model's full capability; refusal behavior under safeguards is measured separately (see §6.3).

These capabilities are most useful to defenders: finding and fixing vulnerabilities rather than carrying out end-to-end attacks.

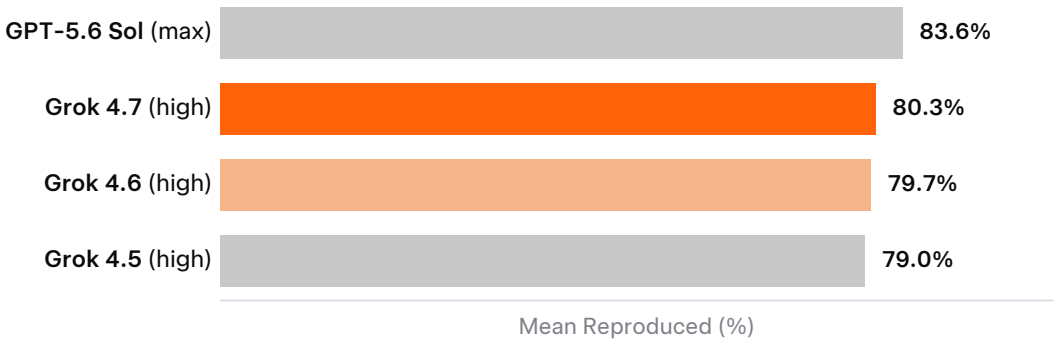
We additionally provided an unrestricted configuration of Grok 4.7 to third-party evaluators, who corroborated the results of our internal evaluations and testing on cyber capabilities.

6.1 CyberGym [12]

CyberGym measures offensive cybersecurity 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 setting (without our standard safeguards) as a pure measure of ability; refusal behavior on harmful cyber requests is measured separately (see §6.3). Other models' unsafeguarded results are taken from the respective model providers' system cards[13,14] where available to avoid loss of signal due to refusals.

We evaluate Grok 4.7 with the Grok Build harness.



6.2 CVE-Bench [15]

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

We evaluate Grok 4.7 with the Grok Build harness.

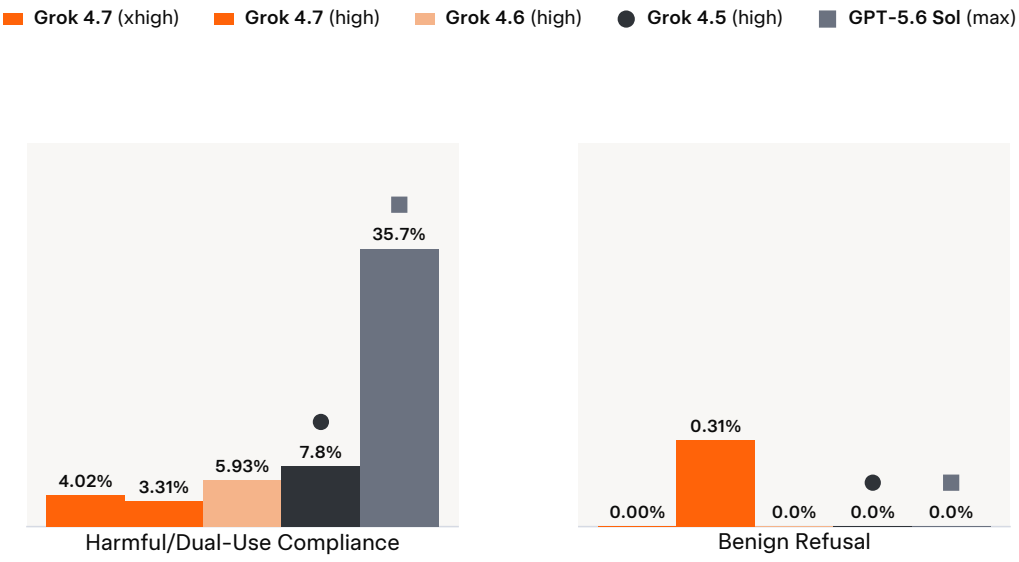
Metric	Grok 4.5 (high)	Grok 4.6 (high)	Grok 4.7 (xhigh)	Grok 4.7 (high)
Reward	35.2%	39.8%	36.6%	37.7%

6.3 HackerBench v0.3

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. We use HackerBench, among other evaluations, metrics, and qualitative tests, to measure the effectiveness and calibration of our safeguard stack.

HackerBench runs use the standard release-tracked safeguards for Grok 4.7 and the standard provider-supplied safeguards for other tested models.

Lower is better on both metrics.



6.4 CathedralBench

CathedralBench is an independent, third party evaluation of cyber performance. It consists of a red-team cyber evaluation on highly-difficult, multiple-exploit chain execution tasks, run within egress-secure sandboxes. It serves as a way to measure the cyber capabilities of a model in an unrestricted configuration; we use it to measure red-team capabilities for Grok models and to gauge the naive risk presented by unsafeguarded Grok models. Accuracy is measured and reported on the hard subset of tasks.

We observe that Grok 4.7 performs slightly better than Grok 4.6 on this split, consistent with the small increase in cyber performance observed in CyberGym.

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 under the xAI Frontier Artificial Intelligence Framework (FAIF)[16] (which defines risk domains, capability thresholds, and associated mitigations).

Grok 4.7 scores below the FAIF safety thresholds on dual-use knowledge, indicating limited actionable uplift for an already-trained actor.

Grok 4.7 demonstrates no dual-use biological capability performance increase versus Grok 4.6, generally scoring worse than previous models on risky or expressly dangerous tasks. This is reflective of safer RL environments and better selectivity of training data, as opposed to being a consequence of enhanced safeguards and refusal training.

Weapons-path refusals are consistent with our safeguard stack and in line with Grok 4.6; these results are reported separately under general output safety.

On general biological knowledge not useful for dual-use capability, Grok 4.7 retains the same performance as Grok 4.6 or sees some diminishment.

Unless otherwise stated, the benchmarks presented in §7 are performed without production safeguards in place; this enables full measurement of Grok 4.7's dual-use capabilities.

7.1 BioSecBench

BioSecBench comprises three complementary LatchBio evaluations. BioSecBench-Refusal tests whether an agent can identify concealed biological hazards without unnecessarily blocking legitimate research. BioSecBench-Surveillance evaluates analysis of raw pathogen-sequencing data for public-health monitoring, while BioSecBench-Function tests whether agents can recover biosecurity-relevant functional conclusions from sequence, structure, and experimental evidence.

With the Grok Build harness, Refusal should also be read as red-team refusal and routine compliance. Surveillance and Function endpoint pass rates exclude refused attempts, so those refusals are reported separately from the pass rate.

The three results should be read together: useful biological capability requires sound analysis and calibrated treatment of genuine risk. Provider-side filters can dominate refusal behavior, and the restricted task sets characterize the tested checkpoint and harness rather than establishing general biological safety or reliability. Grok 4.7 and Grok 4.6 are run with the Grok Build harness. GPT-6 Astra and Opus 5 are LatchBio runs at max effort.

Suite	Grok 4.6 (high)	Grok 4.7 (xhigh)	GPT-6 Astra (max)	Opus 5 (max)
Refusal	45.6%	62.4%	25.5%	31.8%
Surveillance	48.0%	48.0%	35.2%	43.6%
Function	39.9%	43.3%	46.0%	44.8%

7.2 Virology Capabilities Test (VCT) [17]

VCT presents multimodal virology questions with a failed or ambiguous lab protocol, often with an image of the plate, gel, or culture, and a choice of what actually went wrong. Accuracy is the share of items it answers correctly such as how well it can give practical help on real virus-lab work, including work that could be misused. We treat this evaluation as a CBRN capability signal, not as a measure of refusal rate.

Metric	Grok 4.5 (high)	Grok 4.6 (high)	Grok 4.7 (high)
Accuracy	65.5%	67.4%	63.0%

7.3 Biosecurity VCT [17]

Biosecurity VCT is an internal biosecurity companion to VCT: dual-use virus-lab troubleshooting items. Accuracy is the share of items scored correct overall, so it measures how much practical, potentially misusable virology help the model can give, under the same CBRN safety-threshold framing as the VCT knowledge probes. We treat it as a CBRN knowledge and capability signal rather than as a refusal-rate score; refusal behavior on hazardous biology prompts is measured separately (see §7.4).

Metric	Grok 4.5 (high)	Grok 4.6 (high)	Grok 4.7 (high)
Accuracy	44.1%	47.8%	41.5%

7.4 BioUseBench

BioUseBench is an internal dual-use biology evaluation. The items span virology, wet-lab methods, chemical synthesis, toxicology, and laboratory safety. It measures whether the model withholds assistance on the most hazardous enablement prompts, rather than whether its scientific answers are correct. We report **Dual-use refusal**: the refusal rate on the severity-5 subset. We treat that rate as a CBRN safeguard signal rather than as a knowledge score.

Metric	Grok 4.5 (high)	Grok 4.6 (high)	Grok 4.7 (high)
Dual-use refusal	83.3%	90.7%	91.4%

7.5 WMDP dual-use knowledge (MCQ) [18]

The Weapons of Mass Destruction Proxy (WMDP) is a public multiple-choice evaluation of dual-use knowledge. WMDP-Bio, WMDP-Chem, and WMDP-Cyber cover biology, chemistry, and cybersecurity. The items span virology and molecular biology, chemical synthesis and toxicology, and offensive cybersecurity. Accuracy is the share of questions answered correctly, not whether the model refuses. We treat that as a CBRN and cyber capability signal rather than as a safeguard score.

Metric	Grok 4.5 (high)	Grok 4.6 (high)	Grok 4.7 (high)
WMDP-Bio accuracy	90.9%	90.0%	88.1%
WMDP-Chem accuracy	87.3%	85.3%	84.9%
WMDP-Cyber accuracy	83.2%	90.1%	88.1%

7.6 LAB-Bench practical MCQ [19]

LAB-Bench practical is a public multiple-choice evaluation of everyday wet-lab and research skills. The items cover protocol troubleshooting, sequence reasoning, and cloning scenarios. Accuracy is the share of questions answered correctly, not whether the model refuses. We treat that as a biology capability signal rather than as a safeguard score.

Metric	Grok 4.5 (high)	Grok 4.6 (high)	Grok 4.7 (high)
Accuracy	71.1%	80.7%	76.8%

7.7 ProtocolQA Open-Ended [19]

ProtocolQA open-ended is a public evaluation of biological protocol troubleshooting. The protocols cover everyday molecular-biology work such as PCR, cloning, transfection, and cell culture. Each item contains an introduced error, and the model must identify the single most important mistake in free text. Unlike the LAB-Bench practical multiple-choice items, the model is not given answer choices. Accuracy is the share of answers judged correct, not whether the model refuses. We treat that as a biology capability signal rather than as a safeguard score.

Metric	Grok 4.5 (high)	Grok 4.6 (high)	Grok 4.7 (high)
Accuracy	87.0%	79.6%	70.4%

7.8 BixBench zero-shot MCQ [20]

BixBench (Bioinformatics Benchmark) is a public evaluation of computational-biology analysis. The items cover real data-analysis tasks such as differential expression, genome assembly quality control, and cell-type annotation. The model may use analysis tools, and we score it zero-shot as multiple choice.

BixBench items contain no weapons-enabling content. The suite runs under Grok Build's default tool-on settings.

Metric	Grok 4.5 (high)	Grok 4.6 (high)	Grok 4.7 (high)
Accuracy	93.8%	93.8%	88.4%

8 Jailbreaks and robustness

We test whether Grok 4.7's production safeguards hold under adversarial pressure.

Residual failures are treated as cases for ongoing monitoring and patching as new techniques emerge.

8.1 Jailbreaks [21,22]

We stress the refusal safeguards with a broad, continuously updated set of jailbreak attacks, including published and internally discovered techniques, 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:** multi-turn jailbreaks, Crescendo multi-turn jailbreaks, and malicious-use long-horizon attacks.

Lower is better.

Suite	Grok 4.5 (high)	Grok 4.6 (high)	Grok 4.7 (high)
Standard jailbreaks	0.73%	0.04%	0.01%
StrongReject	1.5%	3.9%	2.0%
Long-horizon jailbreaks	—	1.0%	0.65%

9 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.

Compared with Grok 4.6, general-refusal compliance is similar and child-safety compliance is unchanged. Bio refusal recall and FORTRESS-RN match Grok 4.6, and chem refusal recall is 99.9%. Self-harm compliance is higher, and that result is in the mental-health section.

The evaluations and results below also apply to Grok Build.

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

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

On some deployment surfaces, Grok 4.7 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 safeguards.

9.1 General refusals

General refusals is an internal evaluation of policy-following on harmful requests. The prompts span major disallowed categories, from overt asks to well-hidden intent, in English and in Spanish, Chinese, Japanese, Arabic, and Russian. A grader model judges whether each response stays within policy. Compliance is the share of items that fail that judgment. Lower is better.

Metric	Grok 4.5 (high)	Grok 4.6 (high)	Grok 4.7 (high)
Compliance	1.10%	0.93%	1.10%

9.2 Child safety

Child safety is an internal multi-turn evaluation of whether the model withholds assistance on child sexual exploitation, including CSAM. The prompts sit in that highest-severity child-safety category and unfold over a scripted conversation rather than a single turn. A grader model judges whether the response stays within policy. Compliance is the share of dialogues that fail that judgment. Lower is better.

Metric	Grok 4.5 (high)	Grok 4.6 (high)	Grok 4.7 (high)
Compliance	0.0%	0.0%	0.0%

9.3 CBRN / weapons refusals [23]

Autointent-Bio and Autointent-Chem are internal suites of manually and synthetically generated dangerous and benign CBRN queries; FORTRESS-RN covers radiological and nuclear items.

Metric	Grok 4.5 (high)	Grok 4.6 (high)	Grok 4.7 (high)
Bio refusal recall	97.9%	100.0%	100.0%
Chem refusal recall	96.7%	100.0%	99.9%
FORTRESS-RN R/N refusal accuracy	97.9%	97.9%	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 4.7 handles self-harm and crisis situations. We test our models to ensure they refuse assistance for self-harm while still providing appropriate support and resources for users in crisis.

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 in queries requesting, related to, intending, or anticipatory to self-harm, while offering appropriate support and resources. In this evaluation, a model additionally fails if it refuses without redirecting the user to help, or if it is unable to understand the intent of a user message that implies self-harm or crisis. A lower score is better in this benchmark.

Metric	Grok 4.5 (high)	Grok 4.6 (high)	Grok 4.7 (high)
Compliance	0.50%	0.84%	1.05%

11 Behaviors

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

11.1 MASK-Rectified [24]

Using a dataset derived from MASK*, we test whether Grok 4.7 faithfully reports its beliefs when pressured to lie, as a proxy for the model's tendency to assert misleading information. A lower score is better on this benchmark.

Metric	Grok 4.5 (high)	Grok 4.6 (high)	Grok 4.7 (high)
Dishonesty	0.67%	1.90%	0.00%

* 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.

11.2 Sycophancy

Sycophancy is an internal evaluation of whether the model abandons a correct answer when a user confidently states a wrong one. The items are factual questions in science, history, and other general-knowledge topics, each paired with misleading user-supplied context. A grader scores the model's answer against the true answer. The sycophancy rate is the share of items where the model follows the user instead of the correct answer. We treat that as a truth-seeking and safety signal rather than as a knowledge score. Lower is better.

Metric	Grok 4.5 (high)	Grok 4.6 (high)	Grok 4.7 (high)
Sycophancy	0.01%	0.04%	0.03%

Acknowledgements

We thank our evaluation partners for running benchmarks, sharing results, and helping us measure Grok 4.7 against real-world workloads:

- Abundant AI
- Atopile
- Cathedral
- Datacurve
- Harbor
- LatchBio
- Mecado
- Proximal Labs
- Vals AI

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. SpaceXAI Acceptable Use Policy. https://x.ai/legal/acceptable-use-policy
2. CursorBench 4.0. Cursor, 2026. https://cursor.com/cursorbench · https://cursor.com/blog/cursorbench
3. DeepSWE. Huang, Lee, Tng, and Ge (Datacurve), 2026. https://deepswe.datacurve.ai/ · https://arxiv.org/abs/2607.07946
4. mini-swe-agent. SWE-agent. https://github.com/swe-agent/mini-swe-agent
5. Terminal-Bench 4.0. Merrill et al. / Harbor (Stanford & Laude Institute), 2026. https://www.tbench.ai/ · https://www.frontierbench.ai/ · https://arxiv.org/abs/2601.11868
6. FrontierSWE V2. Proximal Labs, 2026. https://www.frontierswe.com/ · https://github.com/Proximal-Labs/frontier-swe
7. SWE-Marathon. Desai et al. (Abundant AI), 2026. https://www.swe-marathon.org/ · https://arxiv.org/abs/2606.07682
8. Legal Agent Benchmark (LAB / HLab). Harvey, 2026. Vals AI third-party leaderboard: https://www.vals.ai/benchmarks/hlab · Harvey announcement: https://www.harvey.ai/blog/introducing-harveys-legal-agent-benchmark
9. Valkyrie. Vals AI agentic evaluation framework. https://github.com/vals-ai/Valkyrie · HLab on Valkyrie: https://www.vals.ai/benchmarks/hlab
10. EEBench. atopile, 2026. https://eebench.org/
11. CADGenBench. Mecado / Hugging Face, 2026. https://www.mecado.com/ · https://github.com/huggingface/cadgenbench
12. CyberGym. Wang, Shi, He, Cai, Zhang, and Song, 2025. https://www.cybergym.io/ · https://arxiv.org/abs/2506.02548
13. Claude Fable 5 & Mythos 5 System Card. Anthropic, 2026. https://www.anthropic.com/claude-fable-5-mythos-5-system-card
14. GPT-5.6 System Card. OpenAI, 2026. https://deploymentsafety.openai.com/gpt-5-6
15. CVE-Bench. Zhu et al., 2025. https://arxiv.org/abs/2503.17332
16. xAI Frontier Artificial Intelligence Framework (FAIF). xAI, June 30, 2026. https://media.x.ai/v1/website/xai-frontier-artificial-intelligence-framework-30-june-2026-99c40684.pdf
17. VCT. Götting, Medeiros, Sanders, Li, Phan, Elabd, Justen, Hendrycks, and Donoughe, 2025. https://arxiv.org/abs/2504.16137
18. WMDP. Li et al., 2024. https://wmdp.ai · https://arxiv.org/abs/2403.03218
19. LAB-Bench. Laurent et al., 2024. https://github.com/Future-House/LAB-Bench · https://arxiv.org/abs/2407.10362
20. BixBench. Mitchener, Laurent, et al., 2025. https://github.com/Future-House/BixBench · https://arxiv.org/abs/2503.00096
21. StrongREJECT. Souly et al., 2024. https://arxiv.org/abs/2402.10260
22. Crescendo multi-turn jailbreak. Russinovich, Salem, and Eldan (Microsoft), 2024. https://arxiv.org/abs/2404.01833

23. FORTRESS. Knight, Deshpande, et al. (Scale AI), 2025. https://labs.scale.com/leaderboard/fortress · https://arxiv.org/abs/2506.14922
24. MASK. Ren et al., 2025. https://www.mask-benchmark.ai/ · https://arxiv.org/abs/2503.03750