



# Model Card: Grok 4.6

August 12, 2026

Revision: 2026-08-**12**



# Model Card: Grok 4.6

August 12, 2026

Revision: 2026-08-**17**

Contents

|     |                                |    |
|-----|--------------------------------|----|
| 1   | Introduction                   | 3  |
| 1.1 | Overview                       | 3  |
| 1.2 | Model development and training | 4  |
| 2   | Coding capabilities            | 5  |
| 2.1 | CursorBench 3.2                | 5  |
| 2.2 | APEX-SWE                       | 7  |
| 2.3 | FrontierCode v1.1              | 8  |
| 2.4 | DeepSWE v1.1                   | 9  |
| 2.5 | SWE-Marathon v1.1              | 10 |
| 2.6 | Terminal-Bench 3.0             | 11 |
| 3   | Engineering acceleration       | 12 |
| 3.1 | EEBench                        | 12 |
| 3.2 | 3DCodeBench                    | 13 |
| 3.4 | CadGenBench - Generation       | 14 |
| 3.5 | CadBench                       | 14 |
| 4   | Office-use abilities           | 15 |
| 4.1 | AA GDPVal                      | 15 |
| 4.2 | AA Briefcase                   | 16 |
| 4.3 | APEX-Agents                    | 17 |
| 4.4 | Vals Index                     | 18 |
| 4.8 | OfficeQA Pro                   | 19 |
| 4.9 | Harvey Legal Agent Benchmark   | 19 |
| 5   | R&D Enablement                 | 20 |
| 5.1 | SpaceXAI MTS Eval              | 20 |
| 5.2 | InferenceEval                  | 21 |
| 5.3 | KernelBenchInternal            | 22 |

Changelog

August 17, 2026

- Updated section numbering for consistency across the card inventory.
- Added PartBench results and graph.
- Added new DeepSearchQA results.
- Updated KernelBenchInternal to version 1.1, reporting results on more difficult task split.
- Clarity updates to prose and figures.
- Corrected eval results on HackerBench v0.2, Self-harm, MASK, LAB.

|                                                              |           |
|--------------------------------------------------------------|-----------|
| <b>6 Search capabilities and factuality</b>                  | <b>23</b> |
| 6.1 Factuality (Hallucination)                               | 23        |
| 6.2 DeepSearchQA                                             | 23        |
| <b>7 Cyber capabilities and safeguards</b>                   | <b>24</b> |
| 7.1 CyberGym                                                 | 24        |
| 7.2 CVE-Bench                                                | 25        |
| 7.3 SecureCodeReview                                         | 25        |
| 7.4 HackerBench v0.2                                         | 26        |
| <b>8 Biological and chemical capabilities and safeguards</b> | <b>27</b> |
| 8.1 Virology Capabilities Test (VCT)                         | 27        |
| 8.2 Biosecurity VCT                                          | 27        |
| 8.3 BioUseBench                                              | 28        |
| 8.4 WMDP dual-use knowledge (MCQ)                            | 28        |
| 8.5 LAB-Bench practical MCQ                                  | 29        |
| 8.6 ProtocolQA Open-Ended                                    | 29        |
| <b>9 Jailbreaks and robustness</b>                           | <b>29</b> |
| 9.1 Jailbreaks                                               | 29        |
| <b>10 General output safety</b>                              | <b>30</b> |
| 10.1 General refusals                                        | 30        |
| 10.2 Child safety                                            | 31        |
| 10.4 CBRN / weapons refusals                                 | 31        |
| <b>11 Mental health</b>                                      | <b>32</b> |
| 11.1 Self-harm refusals                                      | 32        |
| <b>12 Behaviors</b>                                          | <b>32</b> |
| 12.1 MASK-Rectified                                          | 32        |
| 12.2 Sycophancy                                              | 33        |
| <b>References</b>                                            | <b>34</b> |

|                                      |           |
|--------------------------------------|-----------|
| <b>Contents</b>                      |           |
| <b>1 Introduction</b>                | <b>5</b>  |
| 1.1 Overview                         | 5         |
| 1.2 Model development and training   | 6         |
| <b>2 Coding capabilities</b>         | <b>7</b>  |
| 2.1 CursorBench 3.2                  | 7         |
| 2.2 APEX-SWE                         | 9         |
| 2.3 FrontierCode v1.1                | 10        |
| 2.4 DeepSWE v1.1                     | 11        |
| 2.5 SWE-Marathon v1.1                | 12        |
| 2.6 Terminal-Bench 3.0               | 13        |
| <b>3 Knowledge-work capabilities</b> | <b>14</b> |
| 3.1 AA GDPVal                        | 14        |
| 3.2 AA-Briefcase                     | 15        |
| 3.3 APEX-Agents                      | 16        |
| 3.4 OfficeQA Pro                     | 17        |
| 3.5 Legal Agent Benchmark            | 18        |
| <b>4 Engineering acceleration</b>    | <b>19</b> |
| 4.1 EEBench                          | 19        |
| 4.2 3DCodeBench                      | 20        |
| 4.3 PartBench                        | 21        |
| 4.4 CADGenBench                      | 21        |
| 4.5 CADBench                         | 22        |
| <b>5 R&amp;D enablement</b>          | <b>23</b> |
| 5.1 SpaceXAI MTS Eval                | 23        |
| 5.2 InferenceEval                    | 24        |
| 5.3 KernelBenchInternal v1.1         | 25        |

1 Introduction

Grok 4.6 is the latest release in the 1.5T-scale model family of SpaceXAI<sup>¶</sup>, developed in collaboration with Cursor<sup>†</sup>. 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<sup>‡</sup>, 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.

<sup>¶</sup> SpaceXAI is a doing-business-as (dba) name of XAI LLC. xAI and SpaceXAI may be used interchangeably throughout this card.  
<sup>†</sup> Grok 4.6 received supplemental training on anonymized Cursor workflow data to improve coding and agentic performance.  
<sup>‡</sup> SpaceXAI Acceptable Use Policy: <https://x.ai/legal/acceptable-use-policy>.

|      |                                                     |    |
|------|-----------------------------------------------------|----|
| 6    | Search capabilities and factuality                  | 26 |
| 6.1  | Factuality (Hallucination)                          | 26 |
| 6.2  | DeepSearchQA                                        | 27 |
| 7    | Cyber capabilities and safeguards                   | 28 |
| 7.1  | CyberGym                                            | 28 |
| 7.2  | CVE-Bench                                           | 29 |
| 7.3  | SecureCodeReview                                    | 29 |
| 7.4  | HackerBench v0.2                                    | 30 |
| 8    | Biological and chemical capabilities and safeguards | 31 |
| 8.1  | Virology Capabilities Test (VCT)                    | 31 |
| 8.2  | Biosecurity VCT                                     | 32 |
| 8.3  | BioUseBench                                         | 32 |
| 8.4  | WMDP dual-use knowledge (MCQ)                       | 32 |
| 8.5  | LAB-Bench practical MCQ                             | 33 |
| 8.6  | ProtocolQA Open-Ended                               | 33 |
| 8.7  | BixBench zero-shot MCQ                              | 33 |
| 9    | Jailbreaks and robustness                           | 34 |
| 9.1  | Jailbreaks                                          | 34 |
| 10   | General output safety                               | 35 |
| 10.1 | General refusals                                    | 35 |
| 10.2 | Child safety                                        | 36 |
| 10.3 | CBRN / weapons refusals                             | 36 |
| 11   | Mental health                                       | 37 |
| 11.1 | Self-harm refusals                                  | 37 |
| 12   | Behaviors                                           | 37 |
| 12.1 | MASK-Rectified                                      | 37 |
| 12.2 | Sycophancy                                          | 38 |

- 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

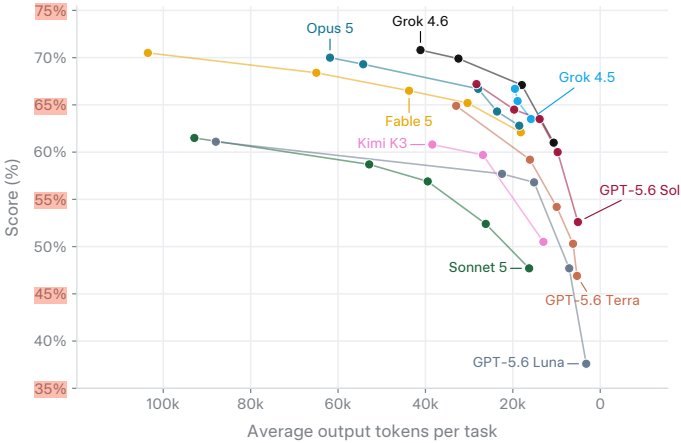
Grok 4.6 is built to be a leading model in coding capabilities, with a strong foundation in software engineering and problem-solving.

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

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>).

1 Introduction

Grok 4.6 is the latest release in SpaceXAI's\* 1.5T-scale model family, developed in collaboration with Cursor<sup>†</sup>. It extends Grok 4.5, demonstrating greater capability and autonomy on coding, engineering, and office work tasks, while also improving in new domains of work (for example, AI research enablement and inference optimization).

It is our most capable model to date.

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.

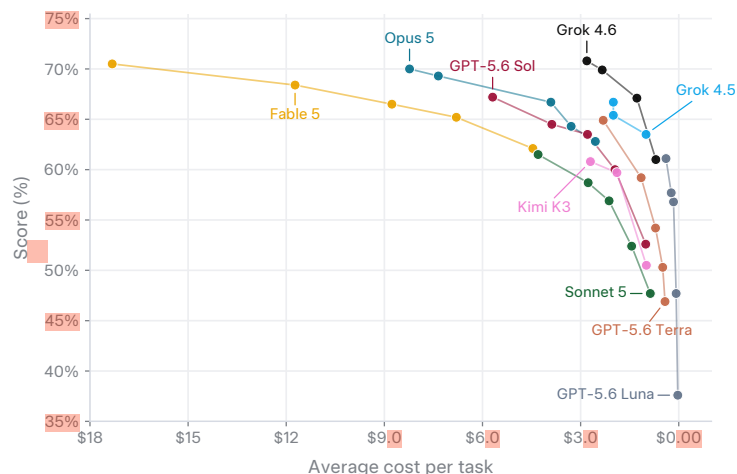
The primary purpose of this model card is to detail the capabilities of Grok 4.6 in quantitative and objective terms, to indicate where this model is most useful. We document safety domains (cyber, bio knowledge, bio agentic, jailbreaks/robustness, general output safety including CBRN refusals, mental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

\* 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.  
‡

## GROK 4.6 MODEL CARD



## GROK 4.6 MODEL CARD

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 [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 4.6 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

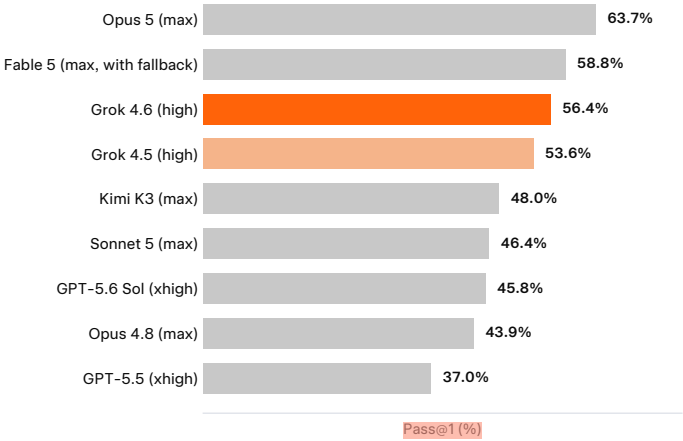
Grok 4.6 was pretrained on publicly available data, data generated internally, as well as other data to which SpaceXAI has secured the necessary rights, followed by supplemental training and post-training with supervised fine-tuning (SFT) and reinforcement learning (RL) on human and synthetic reward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

Supplemental training for Grok 4.6 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 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 4.6 has a pretraining data cutoff of January 2026, and uses data generated as late as June 2026 in its supplemental training.

2.2 APEX-SWE <sup>2</sup>

APEX-SWE measures AI productivity on software-engineering work spanning integration tasks and observability tasks (diagnosing failures from telemetry).<sup>1</sup>



\* Results are taken from tests conducted by Mercor.

2 Coding capabilities

Grok 4.6 is built for strong coding performance across wide-ranging domains of software engineering, problem-solving, and software-engineering-adjacent tasks.

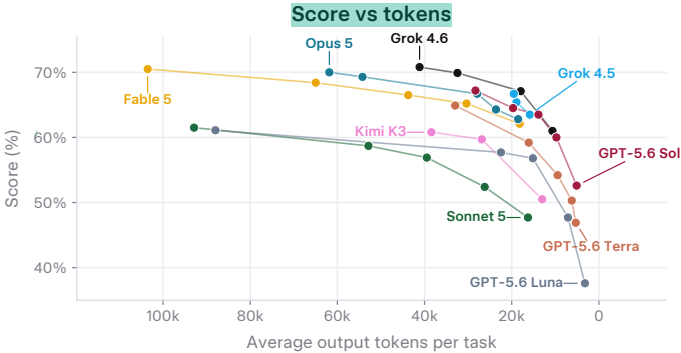
Grok 4.6’s other abilities (in office work, R&D, and CAD) depend on well-developed baseline coding capabilities: an agent that reads, writes, and runs code is a prerequisite for expressive interaction with any digital tool, file, or workflow.

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 3.2 <sup>2</sup>

CursorBench 3.2 evaluates coding agents on realistic IDE-style tasks from production-like Cursor workflows: multi-file edits, natural-language instructions, tools, APIs and MCPs, and iterative fixing.

We evaluate Grok 4.6 with the Cursor agent harness on a fixed set of tasks.\* Grok 4.6 scores 70.8% at xhigh thinking effort, exceeding the other models tested, and 69.9% at high thinking effort, improving on Grok 4.5 high at 66.7%.<sup>1</sup> Averaging 41,136 output tokens per task at xhigh, Grok 4.6 sits at the frontier of capability and efficiency on this benchmark.

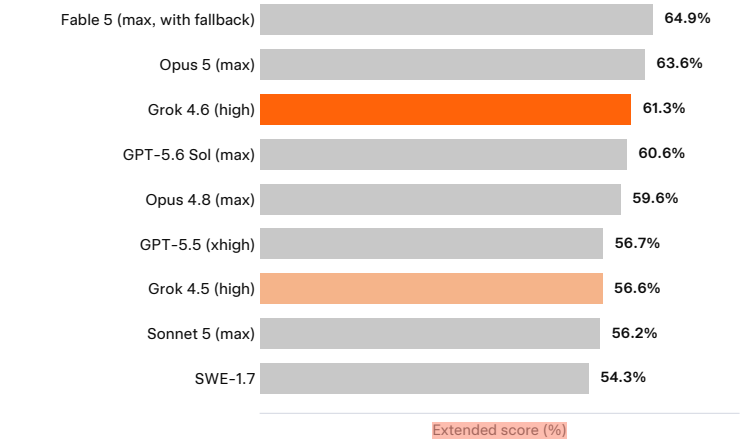


\* Results reported are taken from evaluations conducted by Cursor.  
Grok 4.5 was served with a maximum thinking effort of high.

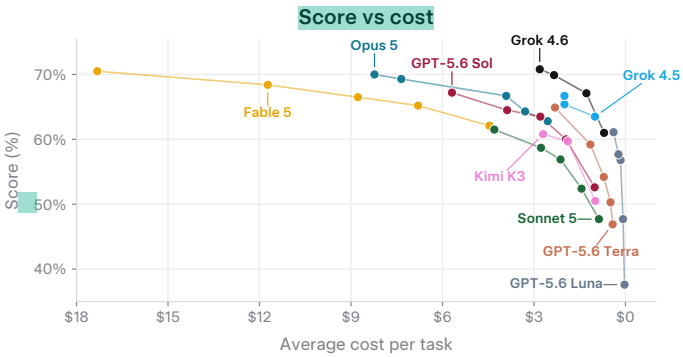
2.3 FrontierCode v1.1

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.



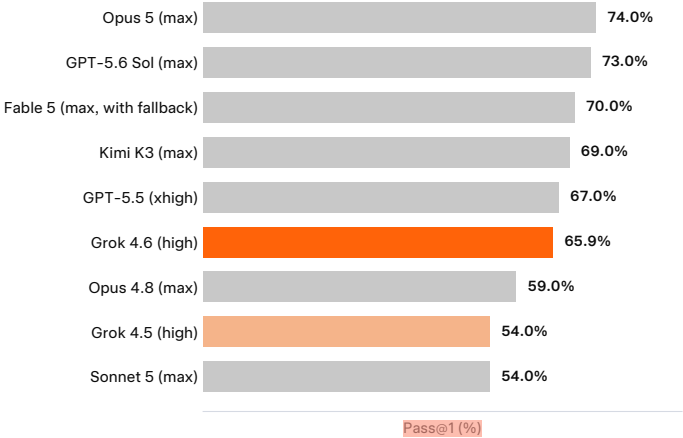
\* Extended set (full private task suite). See <https://cognition.com/frontiercode>.  
SWE-1.7 is plotted without an effort tag. SWE-1.7 is a model from Cognition.



2.4 DeepSWE v1.1

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.

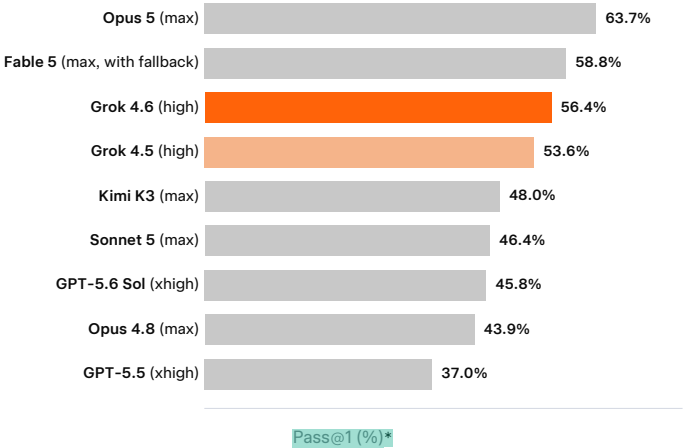


\* mini-swe-agent harness run by Datacurve

2.2 APEX-SWE

APEX-SWE measures AI productivity on software-engineering work spanning integration tasks and observability tasks (e.g. diagnosing failures from telemetry). Coding agents now see active use in maintenance and observability work, representing a materially different task profile from that evaluated by typical software engineering benchmarks. Therefore, we run APEX-SWE to measure Grok 4.6's capabilities in this domain.

Grok 4.6's performance on APEX-SWE is evaluated with the Terminus-2 harness. Grok 4.6 scores 56.4% at high thinking effort.

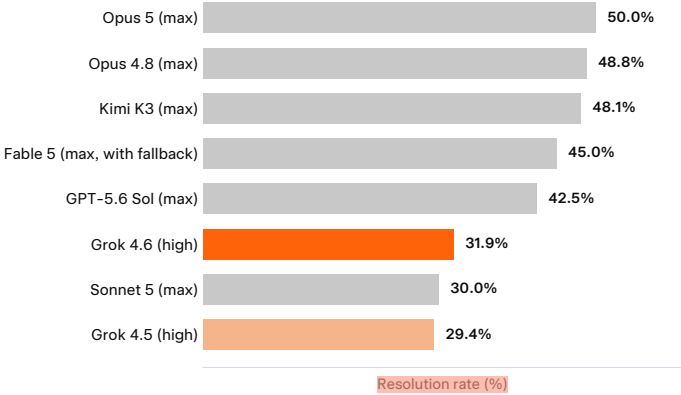


\* Results reported are taken from evaluations conducted by Mercor.

2.5 SWE-Marathon v1.1

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.



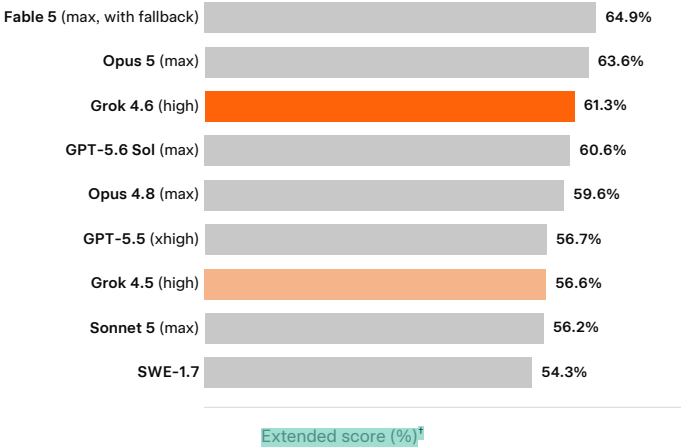
\* Full SWE-Marathon task set: <https://www.swe-marathon.org/>

2.3 FrontierCode v1.1

Correctness alone is an insufficient metric of real-world software engineering utility of a model. Characteristics such as code style, test coverage, and scope-discipline are salient metrics of code utility and quality that are difficult to capture in gold-patch difference benchmarks. FrontierCode v1.1 measures whether AI-generated pull requests would be mergeable by open-source maintainers, measuring behavioral correctness, regression safety, scope discipline, style, and adherence to codebase conventions. We run the extended set of 150 samples to provide a more comprehensive evaluation of the model's capabilities.

Submissions are graded against maintainer-authored rubrics, verifiers, and hard blocker criteria.\*

Grok 4.6's performance on FrontierCode v1.1 is evaluated with the Grok Build harness. Grok 4.6 scores 61.3% at high thinking effort.

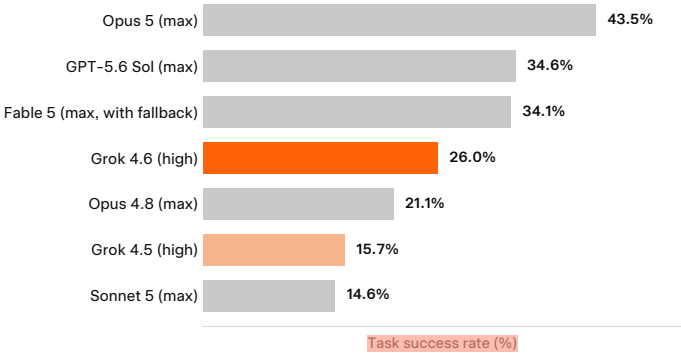


\* SWE-1.7 (Cognition) scores are reported without effort. Results reported are taken from evaluations conducted by Cognition.

2.6 Terminal-Bench 3.0 <sup>6</sup>

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. <sup>6</sup> 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.

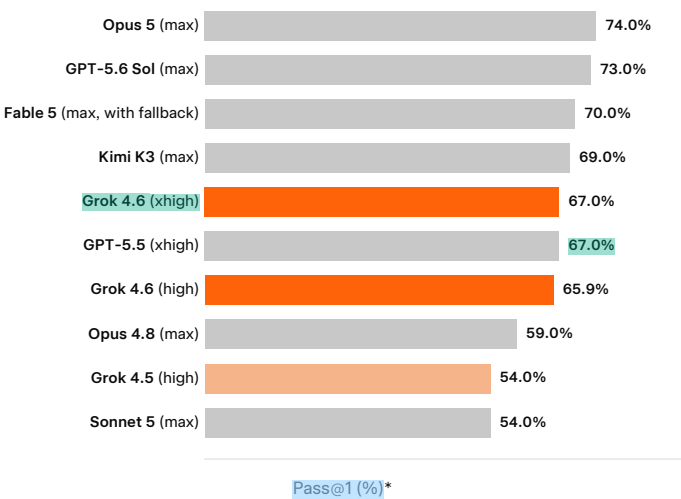


<sup>6</sup> See <https://www.tbench.ai/> and <https://www.frontierbench.ai/>.

2.4 DeepSWE v1.1 <sup>6</sup>

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

Grok 4.6 is evaluated with the mini-swe-agent harness <sup>7</sup>. Grok 4.6 scores 65.9% at high thinking effort and 67.0% at xhigh thinking effort. For non-Grok models, peer numbers use the best result available from model cards or evaluator-reported sources.



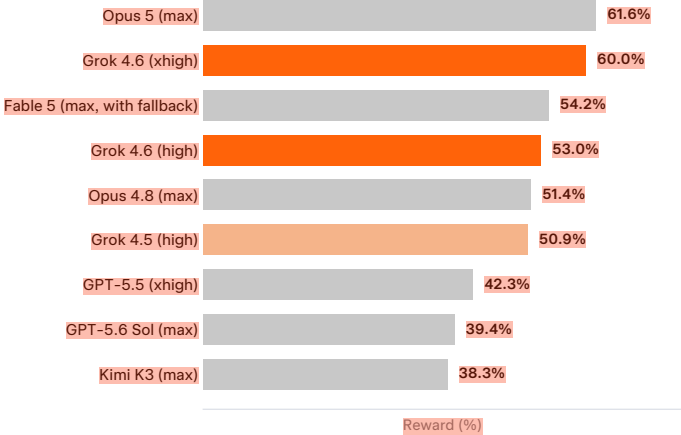
<sup>7</sup> Results reported are taken from evaluations conducted by Datacurve.

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

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



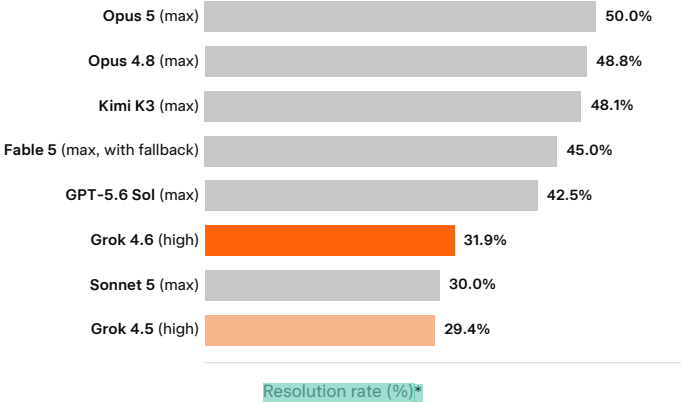
\* Public leaderboard results from <https://eebench.org/> (V1 core corpus).

2.5 SWE-Marathon v1.1

Correctness, goal consistency, and completeness become increasingly important as agents become capable of larger-scope and more open-ended software engineering work.

SWE-Marathon is selected as a benchmark of ultra-long-horizon engineering work that can require millions of tokens and multi-hour trajectories; scoring uses multi-layer verification designed to resist reward hacking and shortcuts.

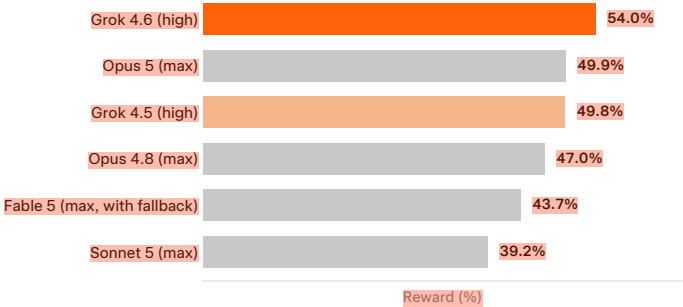
Grok 4.6 is evaluated with the Grok Build harness; peer models use their respective provider harnesses. Grok 4.6 scores 31.9% at high thinking effort.



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

3.2 3DCodeBench <sup>8</sup>

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.\*

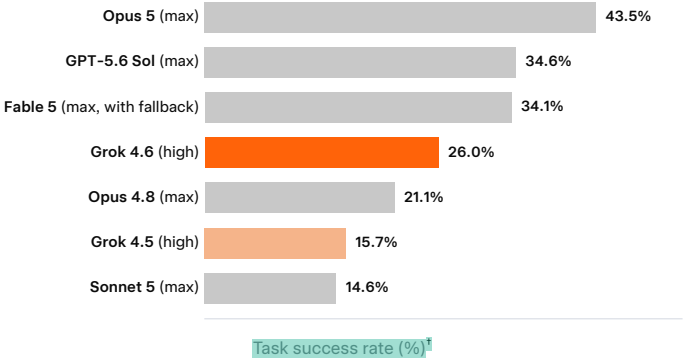


<sup>8</sup> <https://arxiv.org/abs/2606.01057>.

2.6 Terminal-Bench 3.0 <sup>9</sup>

Terminal-Bench 3.0 is the successor [benchmark](#) to Terminal-Bench 2.1,<sup>9</sup> 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.

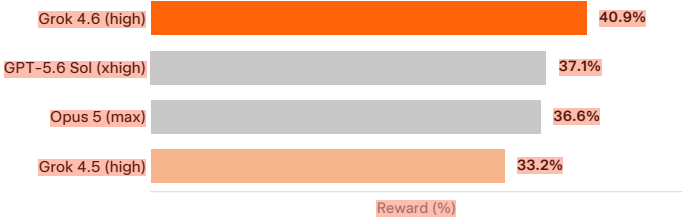
Grok 4.6 is evaluated with the Grok Build harness. Grok 4.6 scores 26.0% at [high thinking effort](#).



<sup>9</sup> The suite was formerly published under the name FrontierBench. Results reported are taken from evaluations conducted by Harbor.

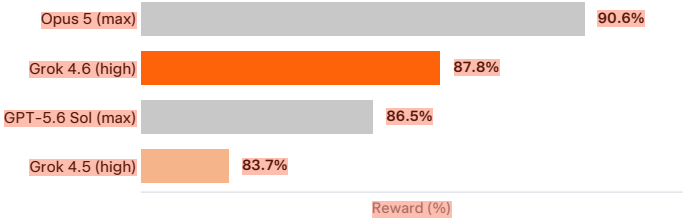
3.4 CadGenBench - Generation <sup>9</sup>

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



3.5 CadBench <sup>10</sup>

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



3 Knowledge-work capabilities

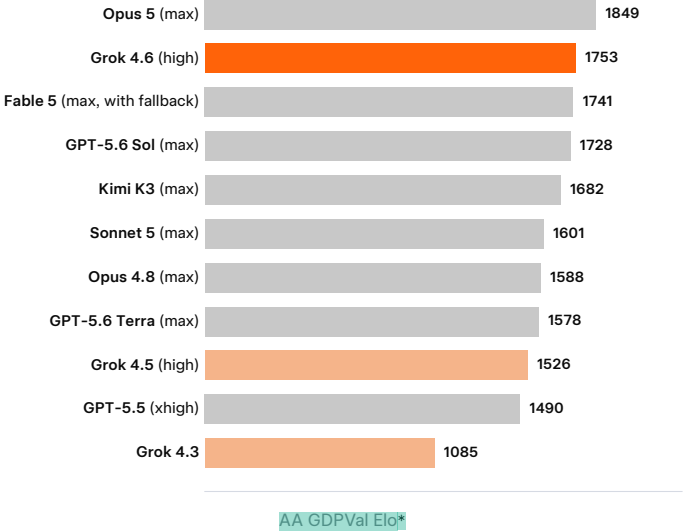
We evaluate performance on professional knowledge-work and structured agent tasks relevant to office and productivity use cases, in order to quantify Grok 4.6’s utility in assisting, accelerating, and automating knowledge work. We select these benchmarks for task- and domain-diversity, as well as for 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 4.6 demonstrates frontier or near-frontier performance.

3.1 AA GDPVal <sup>10</sup>

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. Grok 4.6 scores 1753 Elo at high thinking effort.



\* Results reported are taken from evaluations conducted by Artificial Analysis.

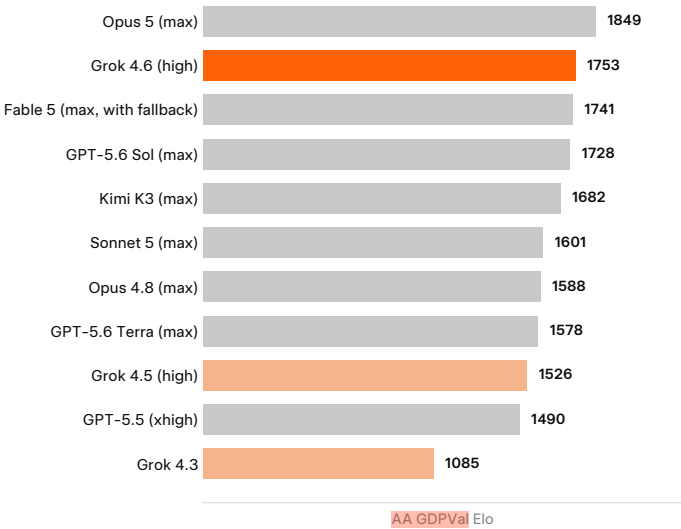
4 Office-use abilities

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

4.1 AA GDPVal <sup>11</sup>

AA GDPVal (GDPval-AA v2) <sup>11</sup> 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.



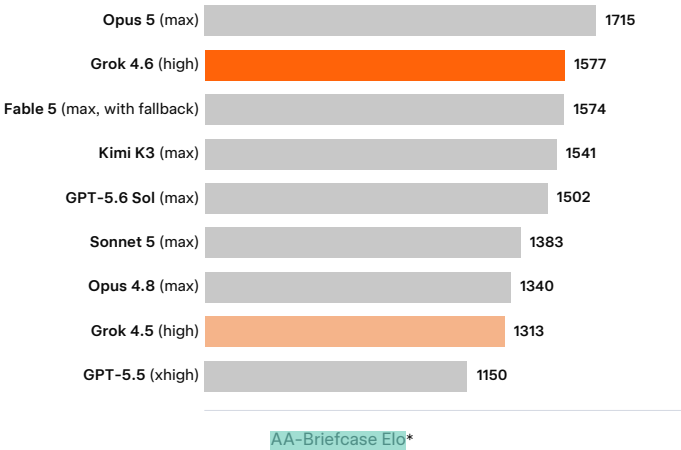
<sup>11</sup> Run by Artificial Analysis (AA) in their harness.

3.2 AA-Briefcase <sup>11</sup>

AA-Briefcase 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. Grok 4.6 is evaluated with the Artificial Analysis harness.

Grok 4.6 scores 1577 Elo at high thinking effort.

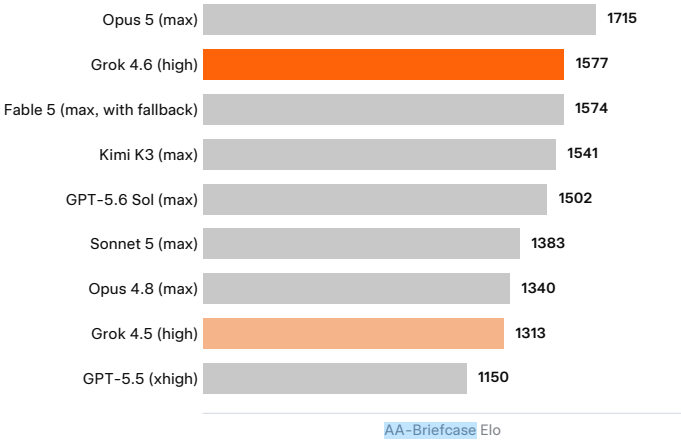


<sup>11</sup> Results reported are taken from evaluations conducted by Artificial Analysis.

4.2 AA Briefcase <sup>12</sup>

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.

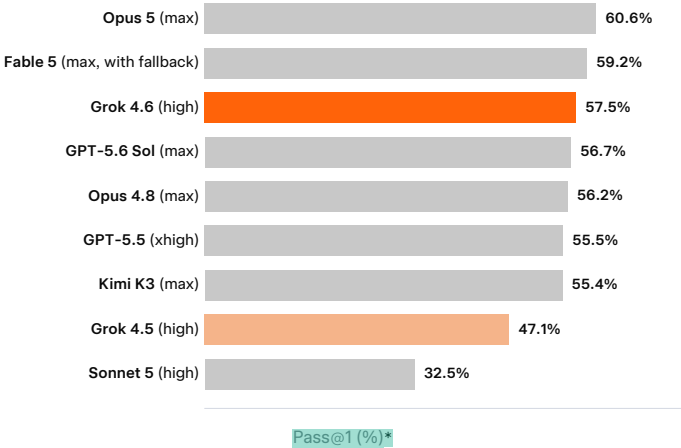


\* <https://artificialanalysis.ai/evaluations/aa-briefcase>

3.3 APEX-Agents <sup>12</sup>

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. Grok 4.6 is evaluated with the Mercor harness. Grok 4.6 scores 57.5% at high thinking effort.

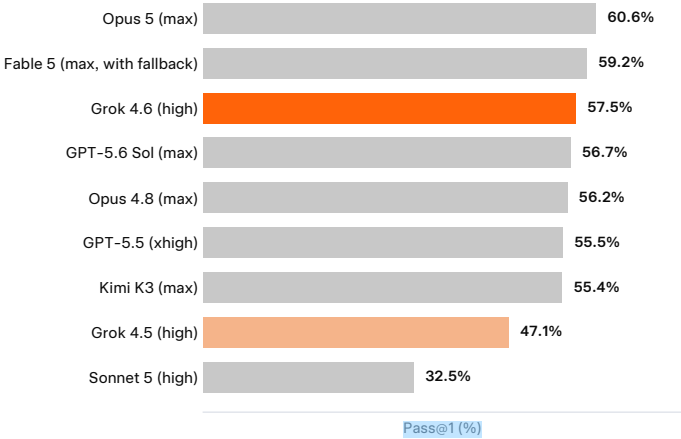


\* Results reported are taken from evaluations conducted by Mercor.

4.3 APEX-Agents <sup>13</sup>

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.

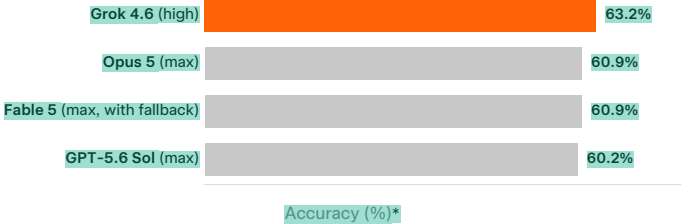


\* <https://www.mercor.com/apex/apex-agents-leaderboard/>

3.4 OfficeQA Pro <sup>13</sup>

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

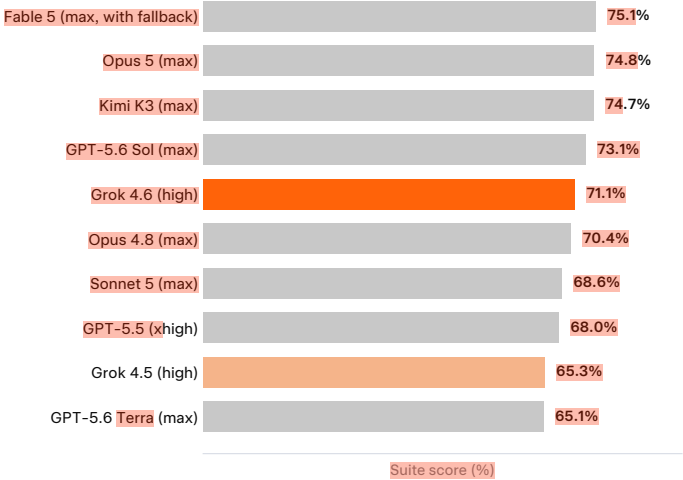
Grok 4.6 is evaluated with the Databricks harness. Grok 4.6 scores 63.2% at high thinking effort.



\* Results are taken from evaluations conducted by Databricks.

4.4 Vals Index<sup>14</sup>

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

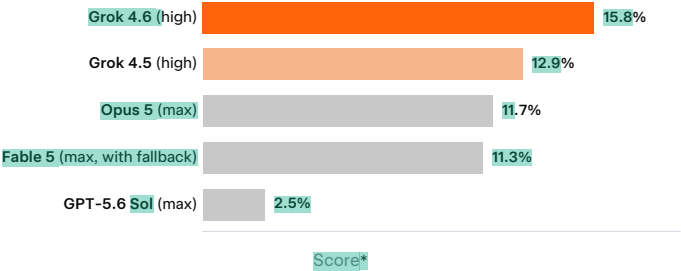


3.5 Legal Agent Benchmark<sup>14</sup>

Legal Agent Benchmark (LAB, also known as the Harvey Legal Agent Benchmark) is a long-horizon legal-agent evaluation in which 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 third-party implementation using the Valkyrie harness<sup>15</sup> (Vals AI's agentic evaluation framework), with a fixed tool and skill set. The agent has six tools: Read File, Edit File, Write File, Glob, Bash, and Grep, as well as three skills: (docx, pptx, and xlsx). The benchmark was adapted to run on Valkyrie and Vals's model library; these are infrastructure changes that do not affect model performance. Harvey's original generation and grading protocol is followed (internet access disabled), and results are reported with the same methodology as Harvey's initial leaderboard.

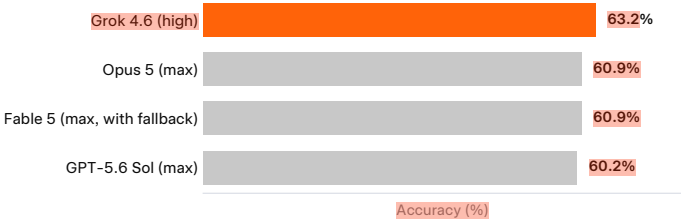
Grok 4.6 is evaluated with the Vals AI (Valkyrie) harness. Grok 4.6 scores 15.8% at high thinking effort.



<sup>14</sup> Results are taken from evaluations conducted by Vals AI.

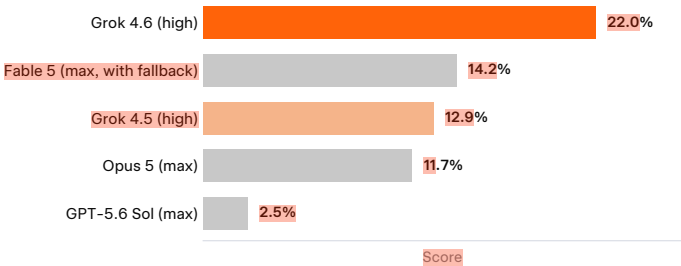
4.8 OfficeQA Pro <sup>15</sup>

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



4.9 Harvey Legal Agent Benchmark <sup>16</sup>

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.



4 Engineering acceleration

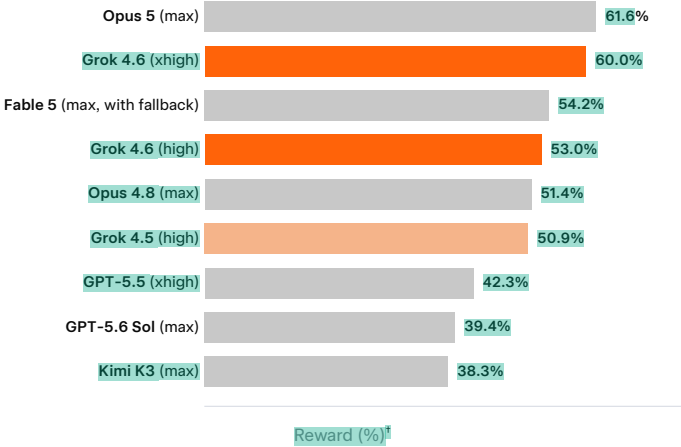
Engineering represents one of the core domains where improved agent capabilities directly support the acceleration of technological development: agents that accelerate rocket design, IC layout, and datacenter power-and-cooling optimization compress the timelines of progress across the physical systems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

We evaluate Grok 4.6 on agentic tasks that accelerate physical-world engineering (i.e. all engineering not purely within the software domain): 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.

4.1 EEBench <sup>16</sup>

EEBench is an electrical engineering and chip design benchmark: models design circuits and other hardware that are graded for physical correctness and functionality.\* Grok 4.6 is evaluated with the Grok Build harness; peer models use their respective provider harnesses.

Grok 4.6 scores 60.0% at xhigh thinking effort and 53.0% at high thinking effort.



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

5 R&D Enablement

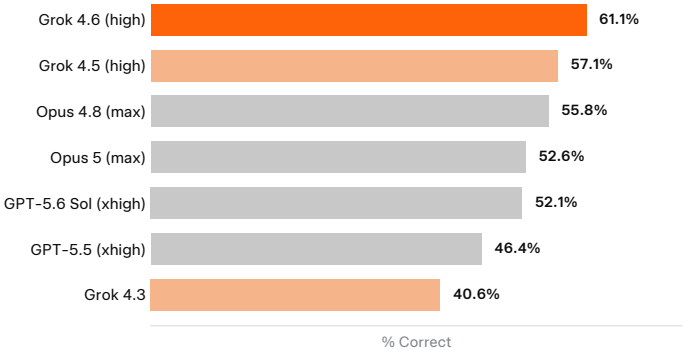
We evaluate Grok 4.6 on its ability to automate parts of the engineering and research process for training and evaluating new versions of itself.

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.\*†



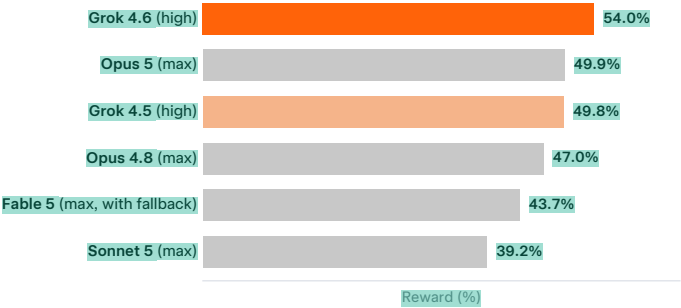
\* GPT-5.6 Sol refused 2 of the 29 tasks and GPT-5.5 refused 1; scores for those models exclude the refused tasks.

† Results were obtained with the Grok Build harness.

4.2 3DCodeBench 17

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. We evaluate Grok 4.6 with the Grok Build harness.

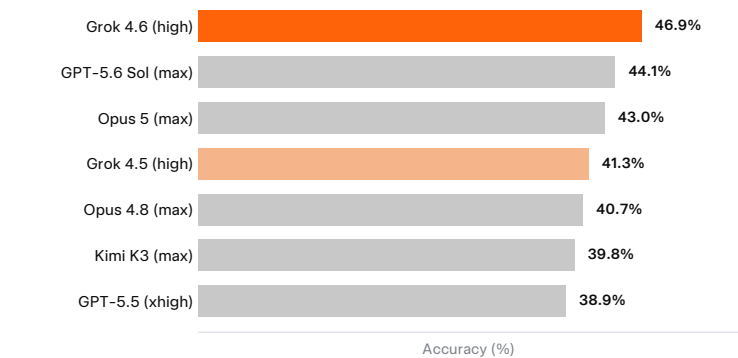
Grok 4.6 scores 54.0% at high thinking effort.



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.



4.3 PartBench

PartBench evaluates agents on parametric mechanical part design and CAD-style generation tasks graded for geometric and manufacturing fidelity.

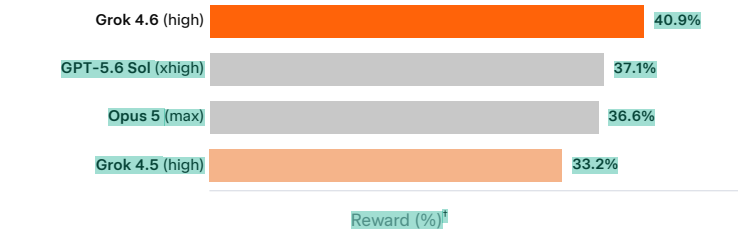
Grok 4.6 is evaluated with the Grok Build harness; peer models use their respective provider harnesses. Grok 4.6 scores 58.5% at xhigh thinking effort.



4.4 CADGenBench

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 4.6 is evaluated with the Grok Build harness; peer models use their respective provider harnesses. Grok 4.6 scores 40.9% at high thinking effort.

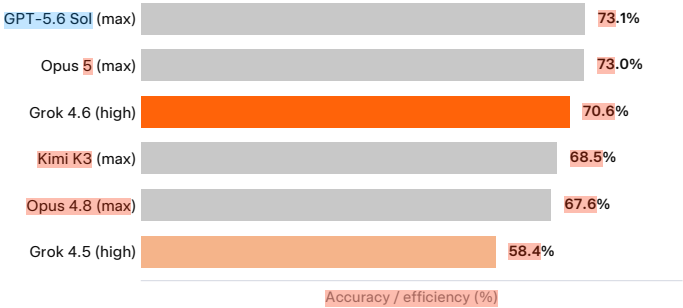


Results are taken from evaluations conducted by Normal Factory.  
Results are taken from evaluations conducted by Mecado.

5.3 KernelBenchInternal

KernelBenchInternal is an internal acceleration suite inspired by KernelBench<sup>17</sup> 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.

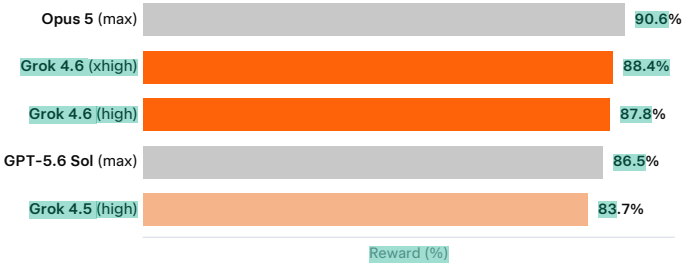
Submitted kernels must compile, match the reference on a correctness trial, and then be timed against a TF32 PyTorch baseline. The reported score depends on both correctness and the size of the measured speedup.



4.5 CADBench<sup>20</sup>

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

Results are taken from the Grok Build harness, run by gNucleus. Grok 4.6 scores 88.4% at **xhigh** thinking effort and 87.8% at **high** thinking effort.\*



\* Additional testing results with different harnesses are reported at <https://cadbench.ai/>.

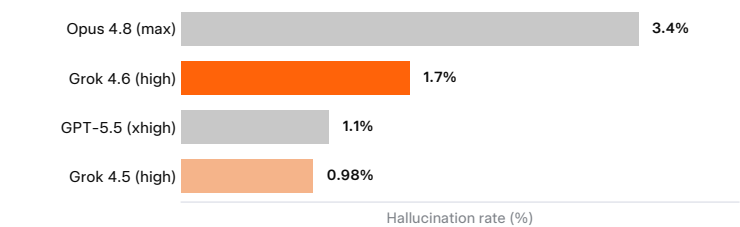
6 Search capabilities and factuality

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

6.1 Factuality (Hallucination)

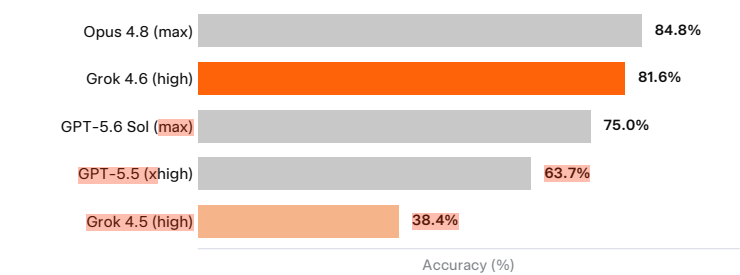
Factuality (Hallucination) measures how often the model asserts unsupported or fabricated claims in a single response to an information-seeking query. Lower is better.

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



6.2 DeepSearchQA

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.

5 R&D enablement

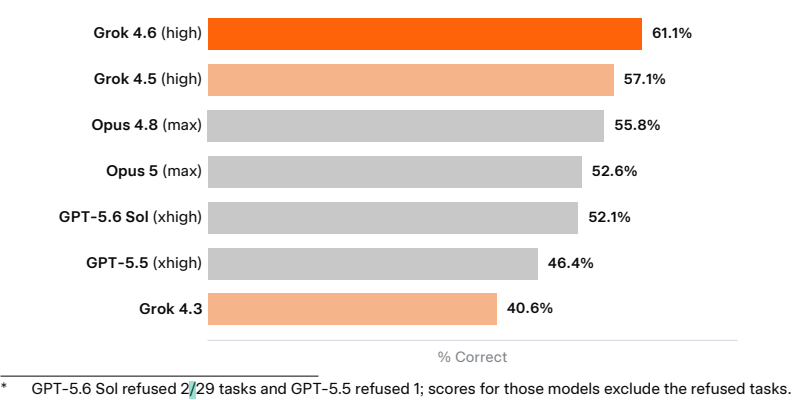
We evaluate Grok 4.6 on its ability to automate parts of the engineering and research process for training and evaluating new versions of itself. Models that can contribute to their own training pipelines, evaluation harnesses, and experimental design enable the development of more capable models, enabling further reinforcement of the model-training loop. We anticipate progress in this domain to significantly contribute to capability improvements in other domains of work, such as engineering, coding, and computer-use; we therefore quantify our model's performance on tasks that measure progress toward this goal.

As a test of autonomous AI-development acceleration, an earlier Grok 4.6 checkpoint was tasked with speeding up its own chat inference, being free to experiment but required to verify end-to-end gains before opening a pull request. In five real-world 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 common in the daily workflows of 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 of the above eval.

Agents are scored on task reward under a fixed rollout budget (time and tokens). Refusals are noted.\* We evaluate Grok 4.6 with the Grok Build harness. Grok 4.6 scores 61.1% at high thinking effort.



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

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.

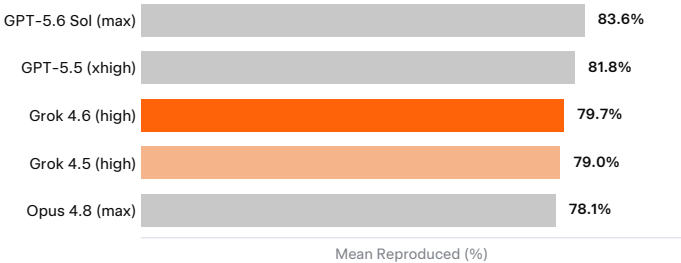
Testing performed by third-party evaluators validate the above results;

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

7.1 CyberGym <sup>19</sup>

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<sup>20,21,22,23</sup>, using the unsafeguarded figures where available to avoid loss of signal due to refusals.

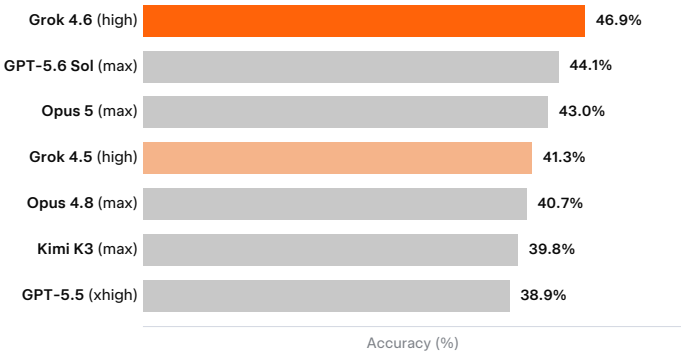


5.2 InferenceEval

InferenceEval is an internal benchmark that tasks agents with implementing real production changes in the code for SpaceXAI's production inference stack. 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.

To more fairly measure capabilities on inference optimization tasks, we intentionally exclude verification of results from the loop; 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. We evaluate Grok 4.6 with the Grok Build harness.

Grok 4.6 scores 46.9% at high thinking effort.



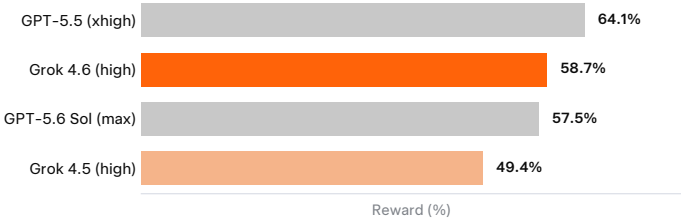
7.2 CVE-Bench 24

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.

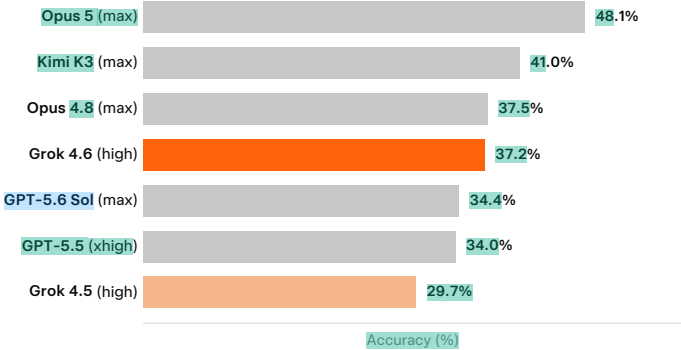


5.3 KernelBenchInternal v1.1

KernelBenchInternal v1.1 is an internal acceleration suite inspired by KernelBench 21 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. We evaluate Grok 4.6 with the Grok Build harness.

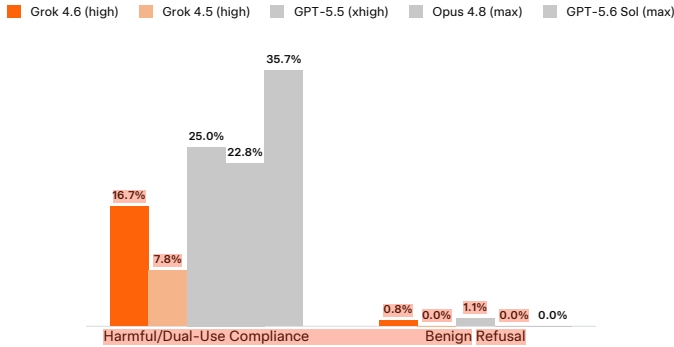
Grok 4.6 scores 37.2% at high thinking effort, improving on Grok 4.5 high at 29.7%.



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.



6 Search capabilities and factuality

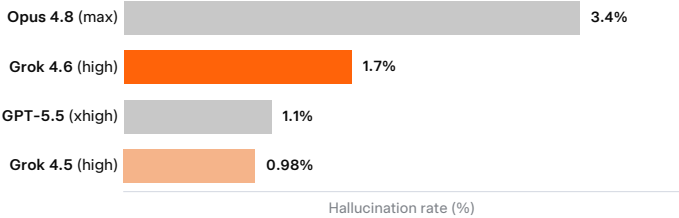
Core to SpaceXAI's mission of truth-seeking is factual correctness, informational recency, and verifiability.

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

6.1 Factuality (Hallucination)

Factuality (Hallucination) is an internal benchmark that 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. We evaluate Grok 4.6 with the Grok Build harness. Grok 4.6 scores 1.7% hallucination rate at high thinking effort.



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)

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

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.

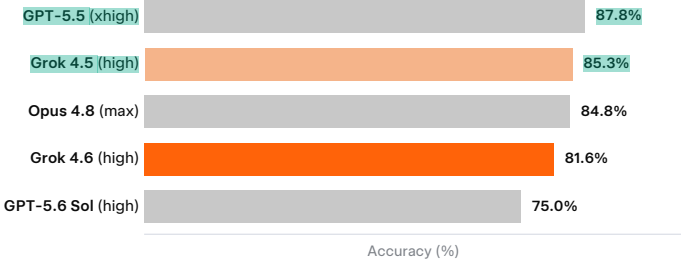
| Metric   | Grok 4.5 (high) | Grok 4.6 (high) |
|----------|-----------------|-----------------|
| Accuracy | —               | 47.8%           |

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

6.2 DeepSearchQA

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.\*

We evaluate Grok 4.6 with the Grok Build harness. Grok 4.6 scores 81.6% at high thinking effort.



\* Evaluation was run on an internal implementation of DeepSearchQA.

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) <sup>26</sup>

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%           |

7 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 4.6 shows some cybersecurity-relevant capability gains over Grok 4.5, concentrated in cyber-defense and vulnerability-mitigation tasks rather than end-to-end offensive 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 §7.4).

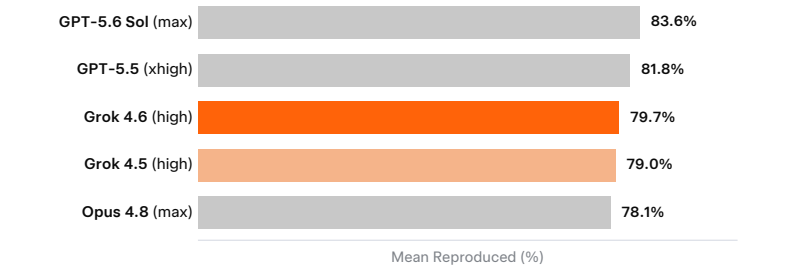
These capabilities are most useful to defenders: finding and fixing vulnerabilities rather than carrying out end-to-end attacks.

We additionally provided an unrestricted configuration of Grok 4.6 to third-party evaluators, who corroborated the results of our internal evaluations and testing on cyber capabilities.

7.1 CyberGym <sup>23</sup>

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 §7.4). Other models' unsafeguarded results are taken from the respective model providers' system cards<sup>24,25,26,27</sup> where available to avoid loss of signal due to refusals.

We evaluate Grok 4.6 with the Grok Build harness. Grok 4.6 scores 79.7% at high thinking effort.



8.5 LAB-Bench practical MCQ <sup>27</sup>

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 <sup>27</sup>

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<sup>28,29</sup>, 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.

7.2 CVE-Bench <sup>28</sup>

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.

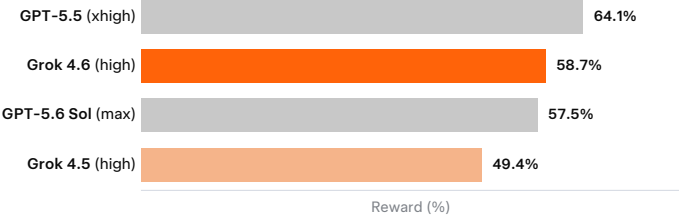
We evaluate Grok 4.6 with the Grok Build harness. Grok 4.6 scores 39.8% at high thinking effort.

| 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.

We evaluate Grok 4.6 with the Grok Build harness. Grok 4.6 scores 58.7% at high thinking effort.



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.

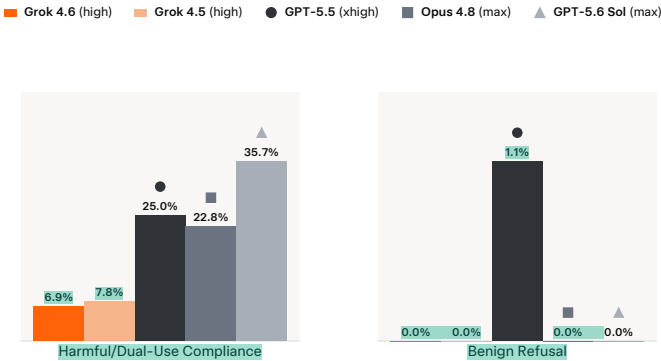
| Metric     | Grok 4.5 (high) | Grok 4.6 (high) |
|------------|-----------------|-----------------|
| Compliance | 1.1%            | 0.93%           |

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. 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 4.6 and the standard provider-supplied safeguards for other tested models.

Grok 4.6 complies with 6.9% of dual-use requests at high thinking effort under our standard safeguard stack. Grok 4.6 has a benign refusal rate of 0.0% at high thinking effort.



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.

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)<sup>29</sup>, 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. Dual-use capability lift versus Grok 4.5 is noted in the biological domain but is limited (for example, a small VCT uptick); weapons-path refusals are consistent with our safeguard stack and show improvement over Grok 4.5; these results are reported separately under general output safety.

On general biological knowledge not useful for dual-use capability, Grok 4.6 shows improvement over Grok 4.5.

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

8.1 Virology Capabilities Test (VCT)<sup>30</sup>

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; a higher score indicates stronger performance on task distributions that overlap with dual-use capability, not better refusal behavior under safeguards.

| Metric   | Grok 4.5 (high) | Grok 4.6 (high) |
|----------|-----------------|-----------------|
| Accuracy | 65.5%           | 67.4%           |

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

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.

8.2 Biosecurity VCT

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 | 44.1%           | 47.8%           |

8.3 BioUseBench

BioUseBench is an internal high-elo dual-use biology suite of biological enablement queries of varying difficulty and risk. We report Dual-use refusal: the refusal rate on the highest-severity (severity-5) dual-use subset. This consists of queries that are non-malicious but could be used in biological uplift work.

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

8.4 WMDP dual-use knowledge (MCQ)

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 | 83.2%           | 90.1%           |

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%           |

8.5 LAB-Bench practical MCQ <sup>32</sup>

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%           |

8.6 ProtocolQA Open-Ended <sup>32</sup>

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%           |

8.7 BixBench zero-shot MCQ <sup>33</sup>

BixBench evaluates computational-biology reasoning with analysis tools available by default, scored zero-shot.

| Metric   | Grok 4.5 (high) | Grok 4.6 (high) |
|----------|-----------------|-----------------|
| Accuracy | 93.8%           | 93.8%           |

BixBench items contain no weapons-enabling content; the suite runs under Grok Build's default tool-on setting.

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/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>

9 Jailbreaks and robustness

We test whether Grok 4.6’s production safeguards hold under adversarial pressure.

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

9.1 Jailbreaks <sup>34,35</sup>

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) |
|-------------------------|-----------------|-----------------|
| Standard jailbreaks     | 0.73%           | 0.04%           |
| StrongReject            | 1.5%            | 3.9%            |
| Long-horizon jailbreaks | —               | 1.0%            |

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>

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, improved CBRN 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 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 4.6 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 safeguards.

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 | 1.10%           | 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%           |

10.3 CBRN / weapons refusals <sup>36</sup>

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) |
|-----------------------------------|-----------------|-----------------|
| Bio refusal recall                | 97.9%           | 100.0%          |
| Chem refusal recall               | 96.7%           | 100.0%          |
| FORTRESS-RN R/N refusal accuracy* | 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.

\* FORTRESS-RN results are taken from evaluations conducted by Scale AI.

11 Mental health

We assess how Grok 4.6 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.

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. 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) |
|------------|-----------------|-----------------|
| Compliance | 0.50%           | 0.84%           |

12 Behaviors

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

12.1 MASK-Rectified <sup>37</sup>

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. A lower score is better in this benchmark.

| Metric     | Grok 4.5 (high) | Grok 4.6 (high) |
|------------|-----------------|-----------------|
| Dishonesty | 0.67%           | 1.90%           |

\* 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%           |

Acknowledgements

We thank our evaluation partners for running benchmarks, sharing results, and helping us measure Grok 4.6 against real-world workloads:

- |                       |             |                  |
|-----------------------|-------------|------------------|
| • Abundant AI         | • Datacurve | • Mercor         |
| • Artificial Analysis | • gNucleus  | • Normal Factory |
| • Atopile             | • Harbor    | • Scale AI       |
| • Cognition           | • Harvey    | • Vals AI        |
| • Databricks          | • Mecado    |                  |

## 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 3.2. Cursor, 2026. <https://cursor.com/cursorbench> · <https://cursor.com/blog/cursorbench>
3. APEX-SWE. Kottamasu et al. (Mercor), 2026. <https://www.mercor.com/apex/apex-swe-leaderboard/> · <https://arxiv.org/abs/2601.08806>
4. Terminus-2. Harbor Framework. <https://www.harborframework.com/docs/agents/terminus-2>
5. FrontierCode v1.1. Cognition, 2026. <https://cognition.com/frontiercode>
6. DeepSWE. Huang, Lee, Tng, and Ge (Datacurve), 2026. <https://deepswe.datacurve.ai/> · <https://arxiv.org/abs/2607.07946>
7. mini-swe-agent. SWE-agent. <https://github.com/swe-agent/mini-swe-agent>
8. SWE-Marathon. Desai et al. (Abundant AI), 2026. <https://www.swe-marathon.org/> · <https://arxiv.org/abs/2606.07682>
9. Terminal-Bench 3.0. Merrill et al. / Harbor (Stanford & Laude Institute), 2026. <https://www.tbench.ai/> · <https://www.frontierbench.ai/> · <https://arxiv.org/abs/2601.11868>
10. AA GDPVal (GDPval-AA v2). Artificial Analysis. <https://artificialanalysis.ai/evaluations/gdpval-aa> · OpenAI GDPval: Patwardhan et al., 2025. <https://arxiv.org/abs/2510.04374>
11. AA-Briefcase. Artificial Analysis, 2026. <https://artificialanalysis.ai/evaluations/aa-briefcase>
12. APEX-Agents. Vidgen et al. (Mercor), 2026. <https://www.mercor.com/apex/apex-agents-leaderboard/> · <https://arxiv.org/abs/2601.14242>
13. OfficeQA Pro. Opsahl-Ong et al. (Databricks), 2026. <https://arxiv.org/abs/2603.08655> · <https://github.com/databricks/officeqa>
14. 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>
15. Valkyrie. Vals AI agentic evaluation framework. <https://github.com/vals-ai/Valkyrie> · HLab on Valkyrie: <https://www.vals.ai/benchmarks/hlab>
16. EEBench. atopile, 2026. <https://eebench.org/>
17. 3DCodeBench. Gao, Shu, Ye, Xiong, Makadia, Guo, Itti, and Chen, 2026. <https://arxiv.org/abs/2606.01057>
18. PartBench. Normal Factory, 2026. <https://www.normalfactory.com/>
19. CADGenBench. Mecado / Hugging Face, 2026. <https://www.mecado.com/> · <https://github.com/huggingface/cadgenbench>
20. CADBench (Parametric CAD Bench). gNucleus. <https://www.gnucleus.ai/cad-bench>
21. KernelBench. Ouyang, Guo, Arora, et al. (Stanford Scaling Intelligence), 2025. <https://arxiv.org/abs/2502.10517> · <https://github.com/ScalingIntelligence/KernelBench>
22. DeepSearchQA. Gupta, Chatterjee, et al., 2026. <https://arxiv.org/abs/2601.20975>
23. CyberGym. Wang, Shi, He, Cai, Zhang, and Song, 2025. <https://www.cybergym.io/> · <https://arxiv.org/abs/2506.02548>

#### GROK 4.6 MODEL CARD

24. Claude Fable 5 & Mythos 5 System Card. Anthropic, 2026. <https://www.anthropic.com/claude-fable-5-mythos-5-system-card>
25. GPT-5.6 System Card. OpenAI, 2026. <https://deploymentsafety.openai.com/gpt-5-6>
26. Claude Opus 4.8 System Card. Anthropic, 2026. <https://www.anthropic.com/claude-opus-4-8-system-card>
27. GPT-5.5 System Card. OpenAI, 2026. <https://deploymentsafety.openai.com/gpt-5-5>
28. CVE-Bench. Zhu et al., 2025. <https://arxiv.org/abs/2503.17332>
29. 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>
30. VCT. Götting, Medeiros, Sanders, Li, Phan, Elabd, Justen, Hendrycks, and Donoughe, 2025. <https://arxiv.org/abs/2504.16137>
31. WMDP. Li et al., 2024. <https://wmdp.ai> · <https://arxiv.org/abs/2403.03218>
32. LAB-Bench. Laurent et al., 2024. <https://github.com/Future-House/LAB-Bench> · <https://arxiv.org/abs/2407.10362>
33. BixBench. Mitchener, Laurent, et al., 2025. <https://github.com/Future-House/BixBench> · <https://arxiv.org/abs/2503.00096>
34. StrongREJECT. Souly et al., 2024. <https://arxiv.org/abs/2402.10260>
35. Crescendo multi-turn jailbreak. Russinovich, Salem, and Eldan (Microsoft), 2024. <https://arxiv.org/abs/2404.01833>
36. FORTRESS. Knight, Deshpande, et al. (Scale AI), 2025. <https://labs.scale.com/leaderboard/fortress> · <https://arxiv.org/abs/2506.14922>
37. MASK. Ren et al., 2025. <https://www.mask-benchmark.ai/> · <https://arxiv.org/abs/2503.03750>

# Comparison Summary Report

card-4p6-4cd2dc57 (1).pdf compared with card-4p6-4cd2dc57.pdf

2026-08-18 20:49:14

596 changes

---

## Text Changes 596

|              |     |
|--------------|-----|
| Insertions   | 177 |
| Deletions    | 135 |
| Replacements | 279 |
| Moves        | 5   |

---

## Visual Changes 0

|              |   |
|--------------|---|
| Insertions   | 0 |
| Deletions    | 0 |
| Replacements | 0 |
| Moves        | 0 |

---

## Formatting Changes 0

---

## Total Changes 596

# Revisions Summary Report

card-4p6-4cd2dc57 (1).pdf compared with card-4p6-4cd2dc57.pdf

2026-08-18 20:49:14

596 changes

Added 177 Removed 135 Replaced 279 Moved 5 Formatting 0

|                 |                 |                                                                                                                                                                                                                                                                                                                                                         |
|-----------------|-----------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| #1 (p.1)        | Replaced (Text) | Before: 12<br>After: 17                                                                                                                                                                                                                                                                                                                                 |
| #2 (p.2)        | Added (Text)    | GROK 4.6 MODEL CARD                                                                                                                                                                                                                                                                                                                                     |
| #3 (p.2)        | Added (Text)    | Changelog                                                                                                                                                                                                                                                                                                                                               |
| #4 (p.2)        | Added (Text)    | August 17, 2026                                                                                                                                                                                                                                                                                                                                         |
| #5 (p.2)        | Added (Text)    | • Updated section numbering for consistency across the card inventory. • Added PartBench results and graph. • Added new DeepSearchQA results. • Updated KernelBenchInternal to version 1.1, reporting results on more difficult task split. • Clarity updates to prose and figures. • Corrected eval results on HackerBench v0.2, Self-harm, MASK, LAB. |
| #6 (p.2)        | Added (Text)    | 1                                                                                                                                                                                                                                                                                                                                                       |
| #7 (p.2 → p.3)  | Replaced (Text) | Before: 3<br>After: 5                                                                                                                                                                                                                                                                                                                                   |
| #8 (p.2 → p.3)  | Replaced (Text) | Before: 3<br>After: 5                                                                                                                                                                                                                                                                                                                                   |
| #9 (p.2 → p.3)  | Replaced (Text) | Before: 4<br>After: 6                                                                                                                                                                                                                                                                                                                                   |
| #10 (p.2 → p.3) | Replaced (Text) | Before: 5<br>After: 7                                                                                                                                                                                                                                                                                                                                   |
| #11 (p.2 → p.3) | Replaced (Text) | Before: 5<br>After: 7                                                                                                                                                                                                                                                                                                                                   |
| #12 (p.2 → p.3) | Replaced (Text) | Before: 7<br>After: 9                                                                                                                                                                                                                                                                                                                                   |
| #13 (p.2 → p.3) | Replaced (Text) | Before: 8<br>After: 10                                                                                                                                                                                                                                                                                                                                  |
| #14 (p.2 → p.3) | Replaced (Text) | Before: 9<br>After: 11                                                                                                                                                                                                                                                                                                                                  |
| #15 (p.2 → p.3) | Replaced (Text) | Before: 10<br>After: 12                                                                                                                                                                                                                                                                                                                                 |
| #16 (p.2 → p.3) | Replaced (Text) | Before: 11<br>After: 13                                                                                                                                                                                                                                                                                                                                 |
| #17 (p.2 → p.3) | Replaced (Text) | Before: Engineering acceleration 12<br>After: Knowledge-work capabilities 14                                                                                                                                                                                                                                                                            |
| #18 (p.2 → p.3) | Replaced (Text) | Before: EEBench .<br>After: AA GDPVal                                                                                                                                                                                                                                                                                                                   |

|                 |                 |                                                                               |
|-----------------|-----------------|-------------------------------------------------------------------------------|
| #19 (p.2 → p.3) | Replaced (Text) | <b>Before:</b> 12<br><b>After:</b> 14                                         |
| #20 (p.2 → p.3) | Replaced (Text) | <b>Before:</b> 13<br><b>After:</b> 15                                         |
| #21 (p.2)       | Removed (Text)  | 4 CadGenBench - Generation                                                    |
| #22 (p.2 → p.3) | Replaced (Text) | <b>Before:</b> 14 3.5 CadBench<br><b>After:</b> . . . . . 16 3.4 OfficeQA Pro |
| #23 (p.2 → p.3) | Replaced (Text) | <b>Before:</b> . . 14<br><b>After:</b> 17 3.5 Legal Agent Benchmark           |
| #24 (p.2)       | Removed (Text)  | 4 Office-use abilities 15 4.1 AA GDPVal                                       |
| #25 (p.2 → p.3) | Replaced (Text) | <b>Before:</b> . . . . . 15 4.2 AA Briefcase<br><b>After:</b> 18              |
| #26 (p.3)       | Added (Text)    | 4 Engineering acceleration 19 4.1 EEBench . .                                 |
| #27 (p.2 → p.3) | Replaced (Text) | <b>Before:</b> 16<br><b>After:</b> 19                                         |
| #28 (p.2 → p.3) | Moved (Text)    | 3 APEX-Agents                                                                 |
| #29 (p.2 → p.3) | Moved (Text)    | 2 3DCodeBench                                                                 |
| #30 (p.2 → p.3) | Replaced (Text) | <b>Before:</b> . 17<br><b>After:</b> 20                                       |
| #31 (p.2 → p.3) | Replaced (Text) | <b>Before:</b> 4 Vals Index<br><b>After:</b> 3 PartBench                      |
| #32 (p.2 → p.3) | Replaced (Text) | <b>Before:</b> 18<br><b>After:</b> 21                                         |
| #33 (p.2 → p.3) | Replaced (Text) | <b>Before:</b> 8 OfficeQA Pro<br><b>After:</b> 4 CADGenBench                  |
| #34 (p.2 → p.3) | Replaced (Text) | <b>Before:</b> . 19<br><b>After:</b> 21                                       |
| #35 (p.2 → p.3) | Replaced (Text) | <b>Before:</b> 9 Harvey Legal Agent Benchmark<br><b>After:</b> 5 CADBench     |
| #36 (p.2 → p.3) | Replaced (Text) | <b>Before:</b> 19<br><b>After:</b> . . . . . 22                               |
| #37 (p.2 → p.3) | Replaced (Text) | <b>Before:</b> E<br><b>After:</b> e                                           |
| #38 (p.2 → p.3) | Replaced (Text) | <b>Before:</b> 20<br><b>After:</b> 23                                         |
| #39 (p.2 → p.3) | Replaced (Text) | <b>Before:</b> 20<br><b>After:</b> 23                                         |
| #40 (p.2 → p.3) | Replaced (Text) | <b>Before:</b> 21<br><b>After:</b> 24                                         |
| #41 (p.2 → p.3) | Replaced (Text) | <b>Before:</b> . .<br><b>After:</b> v1.1                                      |

|                 |                 |                                                                          |
|-----------------|-----------------|--------------------------------------------------------------------------|
| #42 (p.2 → p.3) | Replaced (Text) | Before: 22<br>After: 25                                                  |
| #43 (p.2 → p.3) | Replaced (Text) | Before: 1<br>After: 2                                                    |
| #44 (p.3 → p.4) | Replaced (Text) | Before: 23<br>After: 26                                                  |
| #45 (p.3 → p.4) | Replaced (Text) | Before: 23<br>After: 26                                                  |
| #46 (p.3 → p.4) | Replaced (Text) | Before: 23<br>After: 27                                                  |
| #47 (p.3 → p.4) | Replaced (Text) | Before: 24<br>After: 28                                                  |
| #48 (p.3 → p.4) | Replaced (Text) | Before: 24<br>After: 28                                                  |
| #49 (p.3 → p.4) | Replaced (Text) | Before: 25<br>After: 29                                                  |
| #50 (p.3 → p.4) | Replaced (Text) | Before: 25<br>After: 29                                                  |
| #51 (p.3 → p.4) | Replaced (Text) | Before: 26<br>After: 30                                                  |
| #52 (p.3 → p.4) | Replaced (Text) | Before: 27<br>After: 31                                                  |
| #53 (p.3 → p.4) | Replaced (Text) | Before: 27<br>After: 31                                                  |
| #54 (p.3 → p.4) | Replaced (Text) | Before: 27<br>After: 32                                                  |
| #55 (p.3 → p.4) | Replaced (Text) | Before: 28<br>After: 32                                                  |
| #56 (p.3 → p.4) | Replaced (Text) | Before: 28<br>After: 32                                                  |
| #57 (p.3 → p.4) | Replaced (Text) | Before: 29<br>After: 33                                                  |
| #58 (p.3 → p.4) | Replaced (Text) | Before: 29<br>After: 33 8.7 BixBench zero-shot MCQ . . . . .<br>. . . 33 |
| #59 (p.3 → p.4) | Replaced (Text) | Before: 29<br>After: 34                                                  |
| #60 (p.3 → p.4) | Replaced (Text) | Before: 29<br>After: 34                                                  |
| #61 (p.3 → p.4) | Replaced (Text) | Before: 30<br>After: 35                                                  |
| #62 (p.3 → p.4) | Replaced (Text) | Before: 30<br>After: 35                                                  |

|                 |                 |                                                                                                                                                                                                                                                     |
|-----------------|-----------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| #63 (p.3 → p.4) | Replaced (Text) | <b>Before:</b> 31<br><b>After:</b> 36                                                                                                                                                                                                               |
| #64 (p.3 → p.4) | Replaced (Text) | <b>Before:</b> 4<br><b>After:</b> 3                                                                                                                                                                                                                 |
| #65 (p.3 → p.4) | Replaced (Text) | <b>Before:</b> 31<br><b>After:</b> 36                                                                                                                                                                                                               |
| #66 (p.3 → p.4) | Replaced (Text) | <b>Before:</b> 32<br><b>After:</b> 37                                                                                                                                                                                                               |
| #67 (p.3 → p.4) | Replaced (Text) | <b>Before:</b> 32<br><b>After:</b> 37                                                                                                                                                                                                               |
| #68 (p.3 → p.4) | Replaced (Text) | <b>Before:</b> 32<br><b>After:</b> 37                                                                                                                                                                                                               |
| #69 (p.3 → p.4) | Replaced (Text) | <b>Before:</b> 32<br><b>After:</b> 37                                                                                                                                                                                                               |
| #70 (p.3 → p.4) | Replaced (Text) | <b>Before:</b> 33<br><b>After:</b> 38                                                                                                                                                                                                               |
| #71 (p.3 → p.4) | Replaced (Text) | <b>Before:</b> References 34<br><b>After:</b> 3                                                                                                                                                                                                     |
| #72 (p.3 → p.5) | Replaced (Text) | <b>Before:</b> 2<br><b>After:</b> GROK 4.6 MODEL CARD                                                                                                                                                                                               |
| #73 (p.5)       | Added (Text)    | References 40                                                                                                                                                                                                                                       |
| #74 (p.5)       | Added (Text)    | 4                                                                                                                                                                                                                                                   |
| #75 (p.4 → p.6) | Replaced (Text) | <b>Before:</b> the<br><b>After:</b> SpaceXAI's*                                                                                                                                                                                                     |
| #76 (p.4)       | Removed (Text)  | of SpaceXAI*                                                                                                                                                                                                                                        |
| #77 (p.4 → p.6) | Replaced (Text) | <b>Before:</b> : stronger<br><b>After:</b> , demonstrating greater capability and autonomy                                                                                                                                                          |
| #78 (p.4 → p.6) | Replaced (Text) | <b>Before:</b> on<br><b>After:</b> in                                                                                                                                                                                                               |
| #79 (p.4 → p.6) | Replaced (Text) | <b>Before:</b> or<br><b>After:</b> and                                                                                                                                                                                                              |
| #80 (p.4 → p.6) | Replaced (Text) | <b>Before:</b> strongest<br><b>After:</b> most capable                                                                                                                                                                                              |
| #81 (p.4)       | Removed (Text)  | across the evaluations reported in this card                                                                                                                                                                                                        |
| #82 (p.4 → p.6) | Replaced (Text) | <b>Before:</b> ‡<br><b>After:</b> 1                                                                                                                                                                                                                 |
| #83 (p.6)       | Added (Text)    | * 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. |
| #84 (p.6)       | Added (Text)    | 5                                                                                                                                                                                                                                                   |

|                  |                 |                                                                                                                                                                                                                                                                                                                                                                                         |
|------------------|-----------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| #85 (p.7)        | Added (Text)    | GROK 4.6 MODEL CARD                                                                                                                                                                                                                                                                                                                                                                     |
| #86 (p.7)        | Added (Text)    | Grok by SpaceXAI                                                                                                                                                                                                                                                                                                                                                                        |
| #87 (p.4)        | Removed (Text)  | * 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: <a href="https://x.ai/legal/acceptable-use-policy">https://x.ai/legal/acceptable-use-policy</a> . |
| #88 (p.4)        | Removed (Text)  | 3                                                                                                                                                                                                                                                                                                                                                                                       |
| #89 (p.5)        | Removed (Text)  | GROK 4.6 MODEL CARD                                                                                                                                                                                                                                                                                                                                                                     |
| #90 (p.7)        | Added (Text)    | -                                                                                                                                                                                                                                                                                                                                                                                       |
| #91 (p.7)        | Added (Text)    | for Grok 4.6                                                                                                                                                                                                                                                                                                                                                                            |
| #92 (p.5 → p.7)  | Replaced (Text) | <b>Before:</b> regenerated the<br><b>After:</b> models were used to generate                                                                                                                                                                                                                                                                                                            |
| #93 (p.7)        | Added (Text)    | data                                                                                                                                                                                                                                                                                                                                                                                    |
| #94 (p.7)        | Added (Text)    | , and uses data generated as late as June 2026 in its supplemental training                                                                                                                                                                                                                                                                                                             |
| #95 (p.5 → p.7)  | Replaced (Text) | <b>Before:</b> 4<br><b>After:</b> 6                                                                                                                                                                                                                                                                                                                                                     |
| #96 (p.6 → p.8)  | Replaced (Text) | <b>Before:</b> to be a leading model in<br><b>After:</b> for strong                                                                                                                                                                                                                                                                                                                     |
| #97 (p.6 → p.8)  | Replaced (Text) | <b>Before:</b> capabilities, with a strong foundation in<br><b>After:</b> performance across wide-ranging domains of                                                                                                                                                                                                                                                                    |
| #98 (p.6 → p.8)  | Replaced (Text) | <b>Before:</b> and<br><b>After:</b> ,                                                                                                                                                                                                                                                                                                                                                   |
| #99 (p.8)        | Added (Text)    | , and software-engineering-adjacent tasks                                                                                                                                                                                                                                                                                                                                               |
| #100 (p.6 → p.8) | Replaced (Text) | <b>Before:</b> a strong<br><b>After:</b> well-developed baseline                                                                                                                                                                                                                                                                                                                        |
| #101 (p.6 → p.8) | Replaced (Text) | <b>Before:</b> base<br><b>After:</b> capabilities                                                                                                                                                                                                                                                                                                                                       |
| #102 (p.6 → p.8) | Replaced (Text) | <b>Before:</b> can operate<br><b>After:</b> is a prerequisite for expressive interaction with                                                                                                                                                                                                                                                                                           |
| #103 (p.6 → p.8) | Replaced (Text) | <b>Before:</b> capabilities against<br><b>After:</b> on                                                                                                                                                                                                                                                                                                                                 |
| #104 (p.6 → p.8) | Replaced (Text) | <b>Before:</b> evaluations measuring<br><b>After:</b> benchmarks of                                                                                                                                                                                                                                                                                                                     |
| #105 (p.6 → p.8) | Replaced (Text) | <b>Before:</b> and<br><b>After:</b> ,                                                                                                                                                                                                                                                                                                                                                   |
| #106 (p.6 → p.8) | Replaced (Text) | <b>Before:</b> . Our<br><b>After:</b> , and software-engineering-adjacent tasks, such as SRE and observability. These                                                                                                                                                                                                                                                                   |

|                  |                 |                                                                                                                                                                                                                                                                                                                                                                                                                 |
|------------------|-----------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| #107 (p.6 → p.8) | Replaced (Text) | <b>Before:</b> include<br><b>After:</b> cover                                                                                                                                                                                                                                                                                                                                                                   |
| #108 (p.6)       | Removed (Text)  | program reconstruction,                                                                                                                                                                                                                                                                                                                                                                                         |
| #109 (p.6 → p.8) | Replaced (Text) | <b>Before:</b><br><b>After:</b> -                                                                                                                                                                                                                                                                                                                                                                               |
| #110 (p.6 → p.8) | Replaced (Text) | <b>Before:</b> 1<br><b>After:</b> 2                                                                                                                                                                                                                                                                                                                                                                             |
| #111 (p.6)       | Removed (Text)  | drawn                                                                                                                                                                                                                                                                                                                                                                                                           |
| #112 (p.6 → p.8) | Replaced (Text) | <b>Before:</b> (<br><b>After:</b> :                                                                                                                                                                                                                                                                                                                                                                             |
| #113 (p.8)       | Added (Text)    | natural-language instructions,                                                                                                                                                                                                                                                                                                                                                                                  |
| #114 (p.6 → p.8) | Replaced (Text) | <b>Before:</b> use<br><b>After:</b> s, APIs and MCPs                                                                                                                                                                                                                                                                                                                                                            |
| #115 (p.6 → p.8) | Replaced (Text) | <b>Before:</b> ). *<br><b>After:</b> . We evaluate Grok 4.6 with the Cursor agent harness on a fixed set of tasks.* Grok 4.6 scores 70.8% at xhigh thinking effort, exceeding the other models tested, and 69.9% at high thinking effort, improving on Grok 4.5 high at 66.7%.† Averaging 41,136 output tokens per task at xhigh, Grok 4.6 sits at the frontier of capability and efficiency on this benchmark. |
| #116 (p.6)       | Removed (Text)  | 35%                                                                                                                                                                                                                                                                                                                                                                                                             |
| #117 (p.6)       | Removed (Text)  | 45%                                                                                                                                                                                                                                                                                                                                                                                                             |
| #118 (p.6)       | Removed (Text)  | 55%                                                                                                                                                                                                                                                                                                                                                                                                             |
| #119 (p.6)       | Removed (Text)  | 65%                                                                                                                                                                                                                                                                                                                                                                                                             |
| #120 (p.6)       | Removed (Text)  | 75%                                                                                                                                                                                                                                                                                                                                                                                                             |
| #121 (p.8)       | Added (Text)    |                                                                                                                                                                                                                                                                                                                                                                                                                 |
| #122 (p.8)       | Added (Text)    | Score vs tokens                                                                                                                                                                                                                                                                                                                                                                                                 |
| #123 (p.6 → p.8) | Replaced (Text) | <b>Before:</b> Cursor first-party agent harness (public leaderboard: <a href="https://cursor.com/cursorbench">https://cursor.com/cursorbench</a> ).<br><b>After:</b> Results reported are taken from evaluations conducted by Cursor. † Grok 4.5 was served with a maximum thinking effort of high.                                                                                                             |
| #124 (p.6 → p.8) | Replaced (Text) | <b>Before:</b> 5<br><b>After:</b> 7                                                                                                                                                                                                                                                                                                                                                                             |
| #125 (p.7)       | Removed (Text)  | .0                                                                                                                                                                                                                                                                                                                                                                                                              |
| #126 (p.7)       | Removed (Text)  | .0                                                                                                                                                                                                                                                                                                                                                                                                              |
| #127 (p.7)       | Removed (Text)  | .0                                                                                                                                                                                                                                                                                                                                                                                                              |
| #128 (p.7)       | Removed (Text)  | .00                                                                                                                                                                                                                                                                                                                                                                                                             |
| #129 (p.7)       | Removed (Text)  | 35%                                                                                                                                                                                                                                                                                                                                                                                                             |
| #130 (p.7)       | Removed (Text)  | 45%                                                                                                                                                                                                                                                                                                                                                                                                             |
| #131 (p.7)       | Removed (Text)  | 55%                                                                                                                                                                                                                                                                                                                                                                                                             |

|                   |                 |                                                                                                                                                                                                                                                                                                                                                                                                                                                    |
|-------------------|-----------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| #132 (p.7)        | Removed (Text)  | 65%                                                                                                                                                                                                                                                                                                                                                                                                                                                |
| #133 (p.7 → p.9)  | Replaced (Text) | <b>Before:</b> 75%<br><b>After:</b> Sc                                                                                                                                                                                                                                                                                                                                                                                                             |
| #134 (p.7)        | Removed (Text)  | Sc                                                                                                                                                                                                                                                                                                                                                                                                                                                 |
| #135 (p.9)        | Added (Text)    | Score vs cost                                                                                                                                                                                                                                                                                                                                                                                                                                      |
| #136 (p.7 → p.9)  | Replaced (Text) | <b>Before:</b> 6<br><b>After:</b> 8                                                                                                                                                                                                                                                                                                                                                                                                                |
| #137 (p.8 → p.10) | Replaced (Text) | <b>Before:</b> 2<br><b>After:</b> 3                                                                                                                                                                                                                                                                                                                                                                                                                |
| #138 (p.10)       | Added (Text)    | e.g.                                                                                                                                                                                                                                                                                                                                                                                                                                               |
| #139 (p.8 → p.10) | Replaced (Text) | <b>Before:</b> *<br><b>After:</b> Coding agents now see active use in maintenance and observability work, representing a materially different task profile from that evaluated by typical software engineering benchmarks. Therefore, we run APEX-SWE to measure Grok 4.6's capabilities in this domain. Grok 4.6's performance on APEX-SWE is evaluated with the Terminus-2 harness <sup>4</sup> . Grok 4.6 scores 56.4% at high thinking effort. |
| #140 (p.8)        | Removed (Text)  | Pass@1 (%)                                                                                                                                                                                                                                                                                                                                                                                                                                         |
| #141 (p.10)       | Added (Text)    | Pass@1 (%)*                                                                                                                                                                                                                                                                                                                                                                                                                                        |
| #142 (p.10)       | Added (Text)    | reported                                                                                                                                                                                                                                                                                                                                                                                                                                           |
| #143 (p.8 → p.10) | Replaced (Text) | <b>Before:</b> tests<br><b>After:</b> evaluations                                                                                                                                                                                                                                                                                                                                                                                                  |
| #144 (p.8 → p.10) | Replaced (Text) | <b>Before:</b> 7<br><b>After:</b> 9                                                                                                                                                                                                                                                                                                                                                                                                                |
| #145 (p.9 → p.11) | Replaced (Text) | <b>Before:</b> 3<br><b>After:</b> 5                                                                                                                                                                                                                                                                                                                                                                                                                |
| #146 (p.11)       | Added (Text)    | Correctness alone is an insufficient metric of real-world software engineering utility of a model. Characteristics such as code style, test coverage, and scope-discipline are salient metrics of code utility and quality that are difficult to capture in gold-patch difference benchmarks.                                                                                                                                                      |
| #147 (p.9 → p.11) | Replaced (Text) | <b>Before:</b> :<br><b>After:</b> , measuring                                                                                                                                                                                                                                                                                                                                                                                                      |
| #148 (p.9 → p.11) | Replaced (Text) | <b>Before:</b> *<br><b>After:</b> We run the extended set of 150 samples to provide a more comprehensive evaluation of the model's capabilities.                                                                                                                                                                                                                                                                                                   |
| #149 (p.9 → p.11) | Replaced (Text) | <b>Before:</b> with<br><b>After:</b> against                                                                                                                                                                                                                                                                                                                                                                                                       |
| #150 (p.9 → p.11) | Replaced (Text) | <b>Before:</b> under a fixed agent scaffold. <sup>†</sup><br><b>After:</b> . * Grok 4.6's performance on FrontierCode v1.1 is evaluated with the Grok Build harness. Grok 4.6 scores 61.3% at high thinking effort.                                                                                                                                                                                                                                |
| #151 (p.9)        | Removed (Text)  | Extended score (%)                                                                                                                                                                                                                                                                                                                                                                                                                                 |

|                    |                 |                                                                                                                                                                                                                                                             |
|--------------------|-----------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| #152 (p.9 → p.11)  | Replaced (Text) | <b>Before:</b> * Extended set (full private task suite). See <a href="https://cognition.com/frontiercode">https://cognition.com/frontiercode</a> . † SWE-1.7 is plotted without an effort tag. SWE-1.7 is a model from<br><b>After:</b> Extended score (%)† |
| #153 (p.11)        | Added (Text)    | * SWE-1.7 (Cognition) scores are reported without effort. † Results reported are taken from evaluations conducted by                                                                                                                                        |
| #154 (p.9 → p.11)  | Replaced (Text) | <b>Before:</b> 8<br><b>After:</b> 10                                                                                                                                                                                                                        |
| #155 (p.10 → p.12) | Replaced (Text) | <b>Before:</b> 4<br><b>After:</b> 6                                                                                                                                                                                                                         |
| #156 (p.10)        | Removed (Text)  | *                                                                                                                                                                                                                                                           |
| #157 (p.10 → p.12) | Replaced (Text) | <b>Before:</b> resolution<br><b>After:</b> completion                                                                                                                                                                                                       |
| #158 (p.10 → p.12) | Replaced (Text) | <b>Before:</b> repository issues<br><b>After:</b> long-horizon software engineering tasks                                                                                                                                                                   |
| #159 (p.10 → p.12) | Replaced (Text) | <b>Before:</b> tasks<br><b>After:</b> work                                                                                                                                                                                                                  |
| #160 (p.12)        | Added (Text)    | highly                                                                                                                                                                                                                                                      |
| #161 (p.12)        | Added (Text)    | Grok 4.6 is evaluated with the mini-swe-agent harness <sup>7</sup> . Grok 4.6 scores 65.9% at high thinking effort and 67.0% at xhigh thinking effort.                                                                                                      |
| #162 (p.10)        | Removed (Text)  | model                                                                                                                                                                                                                                                       |
| #163 (p.10)        | Removed (Text)  |                                                                                                                                                                                                                                                             |
| #164 (p.10)        | Removed (Text)  | Pass@1 (%)                                                                                                                                                                                                                                                  |
| #165 (p.12)        | Added (Text)    | Grok 4.6 (xhigh)                                                                                                                                                                                                                                            |
| #166 (p.12)        | Added (Text)    | 67.0%                                                                                                                                                                                                                                                       |
| #167 (p.10 → p.12) | Replaced (Text) | <b>Before:</b> mini-swe-agent harness run<br><b>After:</b> * Results reported are taken from evaluations conducted                                                                                                                                          |
| #168 (p.10 → p.12) | Replaced (Text) | <b>Before:</b><br><b>After:</b> .                                                                                                                                                                                                                           |
| #169 (p.10 → p.12) | Replaced (Text) | <b>Before:</b> 9<br><b>After:</b> 11                                                                                                                                                                                                                        |
| #170 (p.11 → p.13) | Replaced (Text) | <b>Before:</b> 5<br><b>After:</b> 8                                                                                                                                                                                                                         |
| #171 (p.13)        | Added (Text)    | Correctness, goal consistency, and completeness become increasingly important as agents become capable of larger-scope and more open-ended software engineering work.                                                                                       |
| #172 (p.11 → p.13) | Replaced (Text) | <b>Before:</b> * targets<br><b>After:</b> is selected as a benchmark of                                                                                                                                                                                     |
| #173 (p.11 → p.13) | Replaced (Text) | <b>Before:</b> , far beyond a single-PR bug fix. Scoring<br><b>After:</b> ; scoring                                                                                                                                                                         |

|                                    |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                         |
|------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| #174 (p.11 → p.13) Replaced (Text) | <b>Before:</b> Resolution rate (%)<br><b>After:</b> Grok 4.6 is evaluated with the Grok Build harness; peer models use their respective provider harnesses. Grok 4.6 scores 31.9% at high thinking effort.                                                                                                                                                                                                                                                                                                                                                                                                                              |
| #175 (p.11 → p.13) Replaced (Text) | <b>Before:</b> * Full SWE-Marathon task set: <a href="https://www.swe-marathon.org/">https://www.swe-marathon.org/</a> .<br><b>After:</b> Resolution rate (%)*                                                                                                                                                                                                                                                                                                                                                                                                                                                                          |
| #176 (p.11 → p.13) Replaced (Text) | <b>Before:</b> 10<br><b>After:</b> * Results reported are taken from evaluations conducted by Abundant AI.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                              |
| #177 (p.13) Added (Text)           | 12                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                      |
| #178 (p.12 → p.14) Replaced (Text) | <b>Before:</b> 6<br><b>After:</b> 9                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                     |
| #179 (p.14) Added (Text)           | benchmark                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                               |
| #180 (p.12 → p.14) Replaced (Text) | <b>Before:</b> (the suite also published as FrontierBench),<br><b>After:</b> ,*                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                         |
| #181 (p.12) Removed (Text)         | *                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                       |
| #182 (p.12 → p.14) Replaced (Text) | <b>Before:</b> Agents run in a fixed harness on verified tasks spanning long-running commands and multi-step debugging.<br><b>After:</b> Grok 4.6 is evaluated with the Grok Build harness. Grok 4.6 scores 26.0% at high thinking effort.                                                                                                                                                                                                                                                                                                                                                                                              |
| #183 (p.12) Removed (Text)         | Task success rate (%)                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                   |
| #184 (p.12 → p.14) Replaced (Text) | <b>Before:</b> * See <a href="https://www.tbench.ai/">https://www.tbench.ai/</a> and <a href="https://www.frontierbench.ai/">https://www.frontierbench.ai/</a> .<br><b>After:</b> Task success rate (%)†                                                                                                                                                                                                                                                                                                                                                                                                                                |
| #185 (p.12 → p.14) Replaced (Text) | <b>Before:</b> 11<br><b>After:</b> * The suite was formerly published under the name FrontierBench. † Results reported are taken from evaluations conducted by Harbor.                                                                                                                                                                                                                                                                                                                                                                                                                                                                  |
| #186 (p.13 → p.14) Replaced (Text) | <b>Before:</b> GROK 4.6 MODEL CARD<br><b>After:</b> 13                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                  |
| #187 (p.13 → p.15) Replaced (Text) | <b>Before:</b> 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.<br><b>After:</b> GROK 4.6 MODEL CARD |

|                    |                 |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                         |
|--------------------|-----------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| #188 (p.13 → p.15) | Replaced (Text) | <p><b>Before:</b> 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.*</p> <p><b>After:</b> 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 4.6'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 4.6 demonstrates frontier or near-frontier performance.</p> |
| #189 (p.13 → p.15) | Replaced (Text) | <p><b>Before:</b> Reward (%)</p> <p><b>After:</b> 3.1 AA GDPVal 10</p>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                  |
| #190 (p.13)        | Removed (Text)  | Opus 5 (max)                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                            |
| #191 (p.13)        | Removed (Text)  | Grok 4.6 (xhigh)                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                        |
| #192 (p.13)        | Removed (Text)  | Fable 5 (max, with fallback)                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                            |
| #193 (p.13)        | Removed (Text)  | Grok 4.6 (high)                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                         |
| #194 (p.13)        | Removed (Text)  | Opus 4.8 (max)                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                          |
| #195 (p.13)        | Removed (Text)  | Grok 4.5 (high)                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                         |
| #196 (p.13)        | Removed (Text)  | GPT-5.5 (xhigh)                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                         |
| #197 (p.13)        | Removed (Text)  | GPT-5.6 Sol (max)                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                       |
| #198 (p.13)        | Removed (Text)  | Kimi K3 (max)                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                           |
| #199 (p.13)        | Removed (Text)  | 61.6%                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                   |
| #200 (p.13)        | Removed (Text)  | 60.0%                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                   |
| #201 (p.13)        | Removed (Text)  | 54.2%                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                   |
| #202 (p.13)        | Removed (Text)  | 53.0%                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                   |
| #203 (p.13)        | Removed (Text)  | 51.4%                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                   |
| #204 (p.13)        | Removed (Text)  | 50.9%                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                   |
| #205 (p.13)        | Removed (Text)  | 42.3%                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                   |
| #206 (p.13)        | Removed (Text)  | 39.4%                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                   |
| #207 (p.13)        | Removed (Text)  | 38.3%                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                   |
| #208 (p.13)        | Removed (Text)  | * Public leaderboard results from <a href="https://eebench.org/">https://eebench.org/</a> (V1 core corpus).                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                             |
| #209 (p.13)        | Removed (Text)  | 12                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                      |
| #210 (p.14)        | Removed (Text)  | GROK 4.6 MODEL CARD                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                     |
| #211 (p.14)        | Removed (Text)  | <p>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.*</p>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                    |

|             |                |                                                                                                                                                                                   |
|-------------|----------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| #212 (p.14) | Removed (Text) | Reward (%)                                                                                                                                                                        |
| #213 (p.14) | Removed (Text) | Grok 4.6 (high)                                                                                                                                                                   |
| #214 (p.14) | Removed (Text) | Opus 5 (max)                                                                                                                                                                      |
| #215 (p.14) | Removed (Text) | Grok 4.5 (high)                                                                                                                                                                   |
| #216 (p.14) | Removed (Text) | Opus 4.8 (max)                                                                                                                                                                    |
| #217 (p.14) | Removed (Text) | Fable 5 (max, with fallback)                                                                                                                                                      |
| #218 (p.14) | Removed (Text) | Sonnet 5 (max)                                                                                                                                                                    |
| #219 (p.14) | Removed (Text) | 54.0%                                                                                                                                                                             |
| #220 (p.14) | Removed (Text) | 49.9%                                                                                                                                                                             |
| #221 (p.14) | Removed (Text) | 49.8%                                                                                                                                                                             |
| #222 (p.14) | Removed (Text) | 47.0%                                                                                                                                                                             |
| #223 (p.14) | Removed (Text) | 43.7%                                                                                                                                                                             |
| #224 (p.14) | Removed (Text) | 39.2%                                                                                                                                                                             |
| #225 (p.14) | Removed (Text) | * <a href="https://arxiv.org/abs/2606.01057">https://arxiv.org/abs/2606.01057</a> .                                                                                               |
| #226 (p.14) | Removed (Text) | 13                                                                                                                                                                                |
| #227 (p.15) | Removed (Text) | GROK 4.6 MODEL CARD                                                                                                                                                               |
| #228 (p.15) | Removed (Text) | 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. |
| #229 (p.15) | Removed (Text) | Reward (%)                                                                                                                                                                        |
| #230 (p.15) | Removed (Text) | Grok 4.6 (high)                                                                                                                                                                   |
| #231 (p.15) | Removed (Text) | GPT-5.6 Sol (xhigh)                                                                                                                                                               |
| #232 (p.15) | Removed (Text) | Opus 5 (max)                                                                                                                                                                      |
| #233 (p.15) | Removed (Text) | Grok 4.5 (high)                                                                                                                                                                   |
| #234 (p.15) | Removed (Text) | 40.9%                                                                                                                                                                             |
| #235 (p.15) | Removed (Text) | 37.1%                                                                                                                                                                             |
| #236 (p.15) | Removed (Text) | 36.6%                                                                                                                                                                             |
| #237 (p.15) | Removed (Text) | 33.2%                                                                                                                                                                             |
| #238 (p.15) | Removed (Text) | 3.5 CadBench 10 CadBench evaluates agents on parametric CAD design and 3D modeling tasks, including out-of-distribution generation.                                               |
| #239 (p.15) | Removed (Text) | Reward (%)                                                                                                                                                                        |
| #240 (p.15) | Removed (Text) | Opus 5 (max)                                                                                                                                                                      |
| #241 (p.15) | Removed (Text) | Grok 4.6 (high)                                                                                                                                                                   |
| #242 (p.15) | Removed (Text) | GPT-5.6 Sol (max)                                                                                                                                                                 |
| #243 (p.15) | Removed (Text) | Grok 4.5 (high)                                                                                                                                                                   |

|                    |                 |                                                                                                                                                         |
|--------------------|-----------------|---------------------------------------------------------------------------------------------------------------------------------------------------------|
| #244 (p.15)        | Removed (Text)  | 90.6%                                                                                                                                                   |
| #245 (p.15)        | Removed (Text)  | 87.8%                                                                                                                                                   |
| #246 (p.15)        | Removed (Text)  | 86.5%                                                                                                                                                   |
| #247 (p.15)        | Removed (Text)  | 83.7%                                                                                                                                                   |
| #248 (p.15)        | Removed (Text)  | 14                                                                                                                                                      |
| #249 (p.16)        | Removed (Text)  | GROK 4.6 MODEL CARD                                                                                                                                     |
| #250 (p.16)        | Removed (Text)  | 4 Office-use abilities We evaluate performance on professional knowledge-work and structured agent tasks relevant to office and productivity use cases. |
| #251 (p.16)        | Removed (Text)  | 4.1 AA GDPVal 11                                                                                                                                        |
| #252 (p.16)        | Removed (Text)  | *                                                                                                                                                       |
| #253 (p.16 → p.15) | Replaced (Text) | <b>Before:</b> AA GDPVal<br><b>After:</b> Grok 4.6 scores 1753                                                                                          |
| #254 (p.15)        | Added (Text)    | at high thinking effort.                                                                                                                                |
| #255 (p.15)        | Added (Text)    | AA GDPVal Elo*                                                                                                                                          |
| #256 (p.16 → p.15) | Replaced (Text) | <b>Before:</b> Run<br><b>After:</b> Results reported are taken from evaluations conducted                                                               |
| #257 (p.16)        | Removed (Text)  | (AA) in their harness                                                                                                                                   |
| #258 (p.16 → p.15) | Replaced (Text) | <b>Before:</b> 15<br><b>After:</b> 14                                                                                                                   |
| #259 (p.17 → p.16) | Replaced (Text) | <b>Before:</b> 4<br><b>After:</b> 3                                                                                                                     |
| #260 (p.17 → p.16) | Replaced (Text) | <b>Before:</b><br><b>After:</b> -                                                                                                                       |
| #261 (p.17 → p.16) | Replaced (Text) | <b>Before:</b> 12<br><b>After:</b> 11                                                                                                                   |
| #262 (p.17 → p.16) | Replaced (Text) | <b>Before:</b><br><b>After:</b> -                                                                                                                       |
| #263 (p.17)        | Removed (Text)  | (Artificial Analysis)                                                                                                                                   |
| #264 (p.17)        | Removed (Text)  | *                                                                                                                                                       |
| #265 (p.16)        | Added (Text)    | Grok 4.6 is evaluated with the Artificial Analysis harness. Grok 4.6 scores 1577                                                                        |
| #266 (p.17 → p.3)  | Moved (Text)    | AA-Briefcase                                                                                                                                            |
| #267 (p.16)        | Added (Text)    | at high thinking effort.                                                                                                                                |
| #268 (p.16)        | Added (Text)    | AA-Briefcase Elo                                                                                                                                        |
| #269 (p.17 → p.16) | Replaced (Text) | <b>Before:</b> <a href="https://artificialanalysis.ai/">https://artificialanalysis.ai/</a><br><b>After:</b> * Results reported are taken from           |
| #270 (p.17 → p.16) | Replaced (Text) | <b>Before:</b> /aa-briefcase<br><b>After:</b> conducted by Artificial Analysis.                                                                         |

|                    |                 |                                                                                                                                                                                                                                                   |
|--------------------|-----------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| #271 (p.17 → p.16) | Replaced (Text) | <b>Before:</b> 16<br><b>After:</b> 15                                                                                                                                                                                                             |
| #272 (p.18 → p.17) | Replaced (Text) | <b>Before:</b> 4<br><b>After:</b> 3                                                                                                                                                                                                               |
| #273 (p.18 → p.17) | Replaced (Text) | <b>Before:</b> 13<br><b>After:</b> 12                                                                                                                                                                                                             |
| #274 (p.18)        | Removed (Text)  | *                                                                                                                                                                                                                                                 |
| #275 (p.17)        | Added (Text)    | Grok 4.6 is evaluated with the Mercor harness. Grok 4.6 scores 57.5% at high thinking effort.                                                                                                                                                     |
| #276 (p.18 → p.12) | Moved (Text)    | Pass@1 (%)                                                                                                                                                                                                                                        |
| #277 (p.18 → p.17) | Replaced (Text) | <b>Before:</b> * <a href="https://www.mercor.com/apex/apex-agents-leaderboard/">https://www.mercor.com/apex/apex-agents-leaderboard/</a><br><b>After:</b> Pass@1 (%)*                                                                             |
| #278 (p.17)        | Added (Text)    | * Results reported are taken from evaluations conducted by Mercor.                                                                                                                                                                                |
| #279 (p.17)        | Added (Text)    | 16                                                                                                                                                                                                                                                |
| #280 (p.18)        | Added (Text)    | GROK 4.6 MODEL CARD                                                                                                                                                                                                                               |
| #281 (p.18)        | Added (Text)    | 3.4 OfficeQA Pro 13 OfficeQA Pro evaluates models on professional office question-answering across realistic workplace documents and workflows. Grok 4.6 is evaluated with the Databricks harness. Grok 4.6 scores 63.2% at high thinking effort. |
| #282 (p.18)        | Added (Text)    | Grok 4.6 (high)                                                                                                                                                                                                                                   |
| #283 (p.18)        | Added (Text)    | Opus 5 (max)                                                                                                                                                                                                                                      |
| #284 (p.18)        | Added (Text)    | Fable 5 (max, with fallback)                                                                                                                                                                                                                      |
| #285 (p.18)        | Added (Text)    | GPT-5.6 Sol (max)                                                                                                                                                                                                                                 |
| #286 (p.18)        | Added (Text)    | 63.2%                                                                                                                                                                                                                                             |
| #287 (p.18)        | Added (Text)    | 60.9%                                                                                                                                                                                                                                             |
| #288 (p.18)        | Added (Text)    | 60.9%                                                                                                                                                                                                                                             |
| #289 (p.18)        | Added (Text)    | 60.2%                                                                                                                                                                                                                                             |
| #290 (p.18)        | Added (Text)    | Accuracy (%)*                                                                                                                                                                                                                                     |
| #291 (p.18)        | Added (Text)    | * Results are taken from evaluations conducted by Databricks.                                                                                                                                                                                     |
| #292 (p.19)        | Replaced (Text) | <b>Before:</b> 4.4 Vals Index<br><b>After:</b> 3.5 Legal Agent Benchmark                                                                                                                                                                          |

|             |                 |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                    |
|-------------|-----------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| #293 (p.19) | Replaced (Text) | <p><b>Before:</b> Vals Index (Vals AI) is an independent composite of real-world industry and agentic evaluations spanning finance, legal, healthcare, and coding-adjacent professional work.</p> <p><b>After:</b> Legal Agent Benchmark (LAB, also known as the Harvey Legal Agent Benchmark) is a long-horizon legal-agent evaluation in which 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 third-party implementation using the Valkyrie harness<sup>15</sup> (Vals AI's agentic evaluation framework), with a fixed tool and skill set. The agent has six tools: Read File, Edit File, Write File, Glob, Bash, and Grep, as well as three skills: (docx, pptx, and xlsx). The benchmark was adapted to run on Valkyrie and Vals's model library; these are infrastructure changes that do not affect model performance. Harvey's original generation and grading protocol is followed (internet access disabled), and results are reported with the same methodology as Harvey's initial leaderboard. Grok 4.6 is evaluated with the Vals AI (Valkyrie) harness. Grok 4.6 scores 15.8% at high thinking effort.</p> |
| #294 (p.19) | Replaced (Text) | <p><b>Before:</b> Suite score (%)</p> <p><b>After:</b> Grok 4.6 (</p>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                              |
| #295 (p.19) | Removed (Text)  | Fable 5 (max, with fallback)                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                       |
| #296 (p.19) | Removed (Text)  | Opus 5 (max)                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                       |
| #297 (p.19) | Removed (Text)  | Kimi K3 (max)                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                      |
| #298 (p.19) | Removed (Text)  | GPT-5.6 Sol (max)                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                  |
| #299 (p.19) | Removed (Text)  | Grok 4.6 (high)                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                    |
| #300 (p.19) | Removed (Text)  | Opus 4.8 (max)                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                     |
| #301 (p.19) | Removed (Text)  | Sonnet 5 (max)                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                     |
| #302 (p.19) | Removed (Text)  | GPT-5.5 (x                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                         |
| #303 (p.19) | Added (Text)    | Opus 5 (max)                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                       |
| #304 (p.19) | Added (Text)    | Fable 5 (max, with fallback)                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                       |
| #305 (p.19) | Replaced (Text) | <p><b>Before:</b> Terra</p> <p><b>After:</b> Sol</p>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                               |
| #306 (p.19) | Replaced (Text) | <p><b>Before:</b> 75.1</p> <p><b>After:</b> 15.8</p>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                               |
| #307 (p.19) | Replaced (Text) | <p><b>Before:</b> 74.8</p> <p><b>After:</b> 12.9</p>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                               |
| #308 (p.19) | Replaced (Text) | <p><b>Before:</b> 74</p> <p><b>After:</b> 11</p>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                   |
| #309 (p.19) | Replaced (Text) | <p><b>Before:</b> 73.1%</p> <p><b>After:</b> 11.3%</p>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                             |
| #310 (p.19) | Replaced (Text) | <p><b>Before:</b> 71.1%</p> <p><b>After:</b> 2.5%</p>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                              |
| #311 (p.19) | Replaced (Text) | <p><b>Before:</b> 70.4%</p> <p><b>After:</b> Score*</p>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                            |

|             |                 |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                 |
|-------------|-----------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| #312 (p.19) | Replaced (Text) | <b>Before:</b> 68.6%<br><b>After:</b> * Results are taken from evaluations conducted by Vals AI.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                |
| #313 (p.19) | Removed (Text)  | 68.0%                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                           |
| #314 (p.19) | Removed (Text)  | 65.3%                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                           |
| #315 (p.19) | Removed (Text)  | 65.1%                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                           |
| #316 (p.20) | Replaced (Text) | <b>Before:</b> 4.8 OfficeQA Pro 15 OfficeQA Pro evaluates models on professional office question-answering over realistic workplace documents and workflows.<br><b>After:</b> 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 4.6 on agentic tasks that accelerate physical-world engineering (i.e. all engineering not purely within the software domain): 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. |
| #317 (p.20) | Replaced (Text) | <b>Before:</b> Accuracy (%)<br><b>After:</b> 4.1 EEBench 16 EEBench is an electrical engineering and chip design benchmark: models design circuits and other hardware that are graded for physical correctness and functionality.* Grok 4.6 is evaluated with the Grok Build harness; peer models use their respective provider harnesses. Grok 4.6 scores 60.0% at xhigh thinking effort and 53.0% at high thinking effort.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                    |
| #318 (p.20) | Removed (Text)  | Grok 4.6 (high)                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                 |
| #319 (p.20) | Added (Text)    | Grok 4.6 (xhigh)                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                |
| #320 (p.20) | Added (Text)    | Grok 4.6 (high)                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                 |
| #321 (p.20) | Added (Text)    | Opus 4.8 (max)                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                  |
| #322 (p.20) | Added (Text)    | Grok 4.5 (high)                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                 |
| #323 (p.20) | Added (Text)    | GPT-5.5 (xhigh)                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                 |
| #324 (p.20) | Replaced (Text) | <b>Before:</b> 63.2<br><b>After:</b> Kimi K3 (max)                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                              |
| #325 (p.20) | Added (Text)    | 61.6                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                            |
| #326 (p.20) | Replaced (Text) | <b>Before:</b> 60.9%<br><b>After:</b> 60.0%                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                     |
| #327 (p.20) | Replaced (Text) | <b>Before:</b> 60.9%<br><b>After:</b> 54.2%                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                     |
| #328 (p.20) | Replaced (Text) | <b>Before:</b> 60.2%<br><b>After:</b> 53.0%                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                     |

|             |                 |                                                                                                                                                                                                                                                                                                                                                                                 |
|-------------|-----------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| #329 (p.20) | Replaced (Text) | <b>Before:</b> 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.<br><b>After:</b> 51.4% |
| #330 (p.20) | Replaced (Text) | <b>Before:</b> Score<br><b>After:</b> 50.9%                                                                                                                                                                                                                                                                                                                                     |
| #331 (p.20) | Added (Text)    | 42.3%                                                                                                                                                                                                                                                                                                                                                                           |
| #332 (p.20) | Added (Text)    | 39.4%                                                                                                                                                                                                                                                                                                                                                                           |
| #333 (p.20) | Added (Text)    | 38.3%                                                                                                                                                                                                                                                                                                                                                                           |
| #334 (p.20) | Added (Text)    | Reward (%)†                                                                                                                                                                                                                                                                                                                                                                     |
| #335 (p.20) | Added (Text)    | * As part of the V1 core corpus of the benchmark. † Results are taken from evaluations conducted by Atopile.                                                                                                                                                                                                                                                                    |
| #336 (p.20) | Added (Text)    | 19                                                                                                                                                                                                                                                                                                                                                                              |
| #337 (p.21) | Added (Text)    | GROK 4.6 MODEL CARD                                                                                                                                                                                                                                                                                                                                                             |
| #338 (p.21) | Added (Text)    | 4.2 3DCodeBench 17 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. We evaluate Grok 4.6 with the Grok Build harness. Grok 4.6 scores 54.0% at high thinking effort.                                  |
| #339 (p.21) | Added (Text)    | Reward (%)                                                                                                                                                                                                                                                                                                                                                                      |
| #340 (p.21) | Added (Text)    | Grok 4.6 (high)                                                                                                                                                                                                                                                                                                                                                                 |
| #341 (p.21) | Added (Text)    | Opus 5 (max)                                                                                                                                                                                                                                                                                                                                                                    |
| #342 (p.21) | Added (Text)    | Grok 4.5 (high)                                                                                                                                                                                                                                                                                                                                                                 |
| #343 (p.21) | Added (Text)    | Opus 4.8 (max)                                                                                                                                                                                                                                                                                                                                                                  |
| #344 (p.21) | Added (Text)    | Fable 5 (max, with fallback)                                                                                                                                                                                                                                                                                                                                                    |
| #345 (p.21) | Added (Text)    | Sonnet 5 (max)                                                                                                                                                                                                                                                                                                                                                                  |
| #346 (p.21) | Added (Text)    | 54.0%                                                                                                                                                                                                                                                                                                                                                                           |
| #347 (p.21) | Added (Text)    | 49.9%                                                                                                                                                                                                                                                                                                                                                                           |
| #348 (p.21) | Added (Text)    | 49.8%                                                                                                                                                                                                                                                                                                                                                                           |
| #349 (p.21) | Added (Text)    | 47.0%                                                                                                                                                                                                                                                                                                                                                                           |
| #350 (p.21) | Added (Text)    | 43.7%                                                                                                                                                                                                                                                                                                                                                                           |
| #351 (p.21) | Added (Text)    | 39.2%                                                                                                                                                                                                                                                                                                                                                                           |
| #352 (p.21) | Added (Text)    | 20                                                                                                                                                                                                                                                                                                                                                                              |
| #353 (p.22) | Added (Text)    | GROK 4.6 MODEL CARD                                                                                                                                                                                                                                                                                                                                                             |

|                    |                 |                                                                                                                                                                                                                                                                                                                                                                                                |
|--------------------|-----------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| #354 (p.22)        | Added (Text)    | 4.3 PartBench 18 PartBench evaluates agents on parametric mechanical part design and CAD-style generation tasks graded for geometric and manufacturing fidelity. Grok 4.6 is evaluated with the Grok Build harness; peer models use their respective provider harnesses. Grok 4.6 scores 58.5% at xhigh thinking effort.                                                                       |
| #355 (p.22)        | Added (Text)    | Fable 5 (max, with fallback)                                                                                                                                                                                                                                                                                                                                                                   |
| #356 (p.22)        | Added (Text)    | GPT-5.6 Sol (xhigh)                                                                                                                                                                                                                                                                                                                                                                            |
| #357 (p.22)        | Added (Text)    | Grok 4.6 (xhigh)                                                                                                                                                                                                                                                                                                                                                                               |
| #358 (p.22)        | Added (Text)    | Opus 5 (max)                                                                                                                                                                                                                                                                                                                                                                                   |
| #359 (p.22)        | Added (Text)    | 59.9%                                                                                                                                                                                                                                                                                                                                                                                          |
| #360 (p.22)        | Added (Text)    | 59.0%                                                                                                                                                                                                                                                                                                                                                                                          |
| #361 (p.22)        | Added (Text)    | 58.5%                                                                                                                                                                                                                                                                                                                                                                                          |
| #362 (p.22)        | Added (Text)    | 52.4%                                                                                                                                                                                                                                                                                                                                                                                          |
| #363 (p.22)        | Added (Text)    | Reward (%)*                                                                                                                                                                                                                                                                                                                                                                                    |
| #364 (p.22)        | Added (Text)    | 4.4 CADGenBench 19 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 4.6 is evaluated with the Grok Build harness; peer models use their respective provider harnesses. Grok 4.6 scores 40.9% at high thinking effort. |
| #365 (p.20 → p.22) | Replaced (Text) | <b>Before:</b> Fable 5 (max, with fallback)<br><b>After:</b> GPT-5.6 Sol (xhigh)                                                                                                                                                                                                                                                                                                               |
| #366 (p.20 → p.22) | Replaced (Text) | <b>Before:</b> Grok 4.5 (high)<br><b>After:</b> Opus 5 (max)                                                                                                                                                                                                                                                                                                                                   |
| #367 (p.22)        | Added (Text)    | Grok 4.5 (high)                                                                                                                                                                                                                                                                                                                                                                                |
| #368 (p.22)        | Added (Text)    | 40.9%                                                                                                                                                                                                                                                                                                                                                                                          |
| #369 (p.22)        | Added (Text)    | 37.1%                                                                                                                                                                                                                                                                                                                                                                                          |
| #370 (p.22)        | Added (Text)    | 36.6%                                                                                                                                                                                                                                                                                                                                                                                          |
| #371 (p.22)        | Added (Text)    | 33.2%                                                                                                                                                                                                                                                                                                                                                                                          |
| #372 (p.22)        | Added (Text)    | Reward (%)†                                                                                                                                                                                                                                                                                                                                                                                    |
| #373 (p.22)        | Added (Text)    | * Results are taken from evaluations conducted by Normal Factory. † Results are taken from evaluations conducted by Mecado.                                                                                                                                                                                                                                                                    |
| #374 (p.22)        | Added (Text)    | 21                                                                                                                                                                                                                                                                                                                                                                                             |
| #375 (p.23)        | Added (Text)    | GROK 4.6 MODEL CARD                                                                                                                                                                                                                                                                                                                                                                            |
| #376 (p.23)        | Added (Text)    | 4.5 CADBench 20 CADBench evaluates agents on parametric CAD design and 3D modeling tasks, including out-of-distribution generation. Results are taken from the Grok Build harness, run by gNucleus. Grok 4.6 scores 88.4% at xhigh thinking effort and 87.8% at high thinking effort.*                                                                                                         |
| #377 (p.23)        | Added (Text)    | Reward (%)                                                                                                                                                                                                                                                                                                                                                                                     |

|                    |                 |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                          |
|--------------------|-----------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| #378 (p.23)        | Added (Text)    | Grok 4.6 (xhigh)                                                                                                                                                                                                                                                                                                                                                                                                                                                                         |
| #379 (p.23)        | Added (Text)    | Grok 4.6 (high)                                                                                                                                                                                                                                                                                                                                                                                                                                                                          |
| #380 (p.20 → p.23) | Replaced (Text) | <b>Before:</b> 22.0<br><b>After:</b> Grok 4.5 (high)                                                                                                                                                                                                                                                                                                                                                                                                                                     |
| #381 (p.23)        | Added (Text)    | 90.6                                                                                                                                                                                                                                                                                                                                                                                                                                                                                     |
| #382 (p.20 → p.23) | Replaced (Text) | <b>Before:</b> 14.2<br><b>After:</b> 88.4%                                                                                                                                                                                                                                                                                                                                                                                                                                               |
| #383 (p.23)        | Added (Text)    | 87.8                                                                                                                                                                                                                                                                                                                                                                                                                                                                                     |
| #384 (p.20 → p.23) | Replaced (Text) | <b>Before:</b> 12.9<br><b>After:</b> 86.5                                                                                                                                                                                                                                                                                                                                                                                                                                                |
| #385 (p.20 → p.23) | Replaced (Text) | <b>Before:</b> 11<br><b>After:</b> 83                                                                                                                                                                                                                                                                                                                                                                                                                                                    |
| #386 (p.20 → p.23) | Replaced (Text) | <b>Before:</b> 2.5%<br><b>After:</b> * Additional testing results with different harnesses are reported at <a href="https://cadbench.ai/">https://cadbench.ai/</a> .                                                                                                                                                                                                                                                                                                                     |
| #387 (p.20 → p.23) | Replaced (Text) | <b>Before:</b> 19<br><b>After:</b> 22                                                                                                                                                                                                                                                                                                                                                                                                                                                    |
| #388 (p.21 → p.24) | Replaced (Text) | <b>Before:</b> E<br><b>After:</b> e                                                                                                                                                                                                                                                                                                                                                                                                                                                      |
| #389 (p.24)        | Added (Text)    | Models that can contribute to their own training pipelines, evaluation harnesses, and experimental design enable the development of more capable models, enabling further reinforcement of the model-training loop. We anticipate progress in this domain to significantly contribute to capability improvements in other domains of work, such as engineering, coding, and computer-use; we therefore quantify our model's performance on tasks that measure progress toward this goal. |
| #390 (p.21 → p.24) | Replaced (Text) | <b>Before:</b> ML<br><b>After:</b> AI-development                                                                                                                                                                                                                                                                                                                                                                                                                                        |
| #391 (p.24)        | Added (Text)    | being                                                                                                                                                                                                                                                                                                                                                                                                                                                                                    |
| #392 (p.24)        | Added (Text)    | real-world                                                                                                                                                                                                                                                                                                                                                                                                                                                                               |
| #393 (p.21 → p.24) | Replaced (Text) | <b>Before:</b> faced by<br><b>After:</b> common in the daily workflows of                                                                                                                                                                                                                                                                                                                                                                                                                |
| #394 (p.24)        | Added (Text)    | of the above eval                                                                                                                                                                                                                                                                                                                                                                                                                                                                        |
| #395 (p.21 → p.24) | Replaced (Text) | <b>Before:</b> †<br><b>After:</b> We evaluate Grok 4.6 with the Grok Build harness. Grok 4.6 scores 61.1% at high thinking effort.                                                                                                                                                                                                                                                                                                                                                       |
| #396 (p.21 → p.24) | Replaced (Text) | <b>Before:</b> of the<br><b>After:</b> /                                                                                                                                                                                                                                                                                                                                                                                                                                                 |
| #397 (p.21 → p.24) | Replaced (Text) | <b>Before:</b> † Results were obtained with the Grok Build harness.<br><b>After:</b> 23                                                                                                                                                                                                                                                                                                                                                                                                  |
| #398 (p.21)        | Removed (Text)  | 20                                                                                                                                                                                                                                                                                                                                                                                                                                                                                       |
| #399 (p.25)        | Added (Text)    | is an internal benchmark that t                                                                                                                                                                                                                                                                                                                                                                                                                                                          |

|                    |                 |                                                                                                                                                                          |
|--------------------|-----------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| #400 (p.22 → p.25) | Replaced (Text) | <b>Before:</b> to<br><b>After:</b> with                                                                                                                                  |
| #401 (p.25)        | Added (Text)    | ing                                                                                                                                                                      |
| #402 (p.22 → p.25) | Replaced (Text) | <b>Before:</b> that actually runs chat<br><b>After:</b> for SpaceXAI's production                                                                                        |
| #403 (p.25)        | Added (Text)    | stack                                                                                                                                                                    |
| #404 (p.22 → p.25) | Replaced (Text) | <b>Before:</b> S<br><b>After:</b> To more fairly measure capabilities on inference optimization tasks, we intentionally exclude verification of results from the loop; s |
| #405 (p.25)        | Added (Text)    | We evaluate Grok 4.6 with the Grok Build harness. Grok 4.6 scores 46.9% at high thinking effort.                                                                         |
| #406 (p.22 → p.25) | Replaced (Text) | <b>Before:</b> 21<br><b>After:</b> 24                                                                                                                                    |
| #407 (p.26)        | Added (Text)    | v1.1                                                                                                                                                                     |
| #408 (p.26)        | Added (Text)    | v1.1                                                                                                                                                                     |
| #409 (p.23 → p.26) | Replaced (Text) | <b>Before:</b> 17<br><b>After:</b> 21                                                                                                                                    |
| #410 (p.23 → p.26) | Replaced (Text) | <b>Before:</b><br><b>After:</b> We evaluate Grok 4.6 with the Grok Build harness. Grok 4.6 scores 37.2% at high thinking effort, improving on Grok 4.5 high at 29.7%.    |
| #411 (p.23 → p.26) | Replaced (Text) | <b>Before:</b> Accuracy / efficiency (%)<br><b>After:</b> Accuracy (%)                                                                                                   |
| #412 (p.26)        | Added (Text)    | Opus 5 (max)                                                                                                                                                             |
| #413 (p.26)        | Added (Text)    | Kimi K3                                                                                                                                                                  |
| #414 (p.23 → p.26) | Replaced (Text) | <b>Before:</b> 5<br><b>After:</b> 4.8                                                                                                                                    |
| #415 (p.23)        | Removed (Text)  | Kimi K3                                                                                                                                                                  |
| #416 (p.23 → p.26) | Moved (Text)    | GPT-5.6 Sol                                                                                                                                                              |
| #417 (p.23 → p.26) | Replaced (Text) | <b>Before:</b> Opus 4.8 (max<br><b>After:</b> GPT-5.5 (xhigh                                                                                                             |
| #418 (p.23 → p.26) | Replaced (Text) | <b>Before:</b> 73<br><b>After:</b> 48                                                                                                                                    |
| #419 (p.23 → p.26) | Replaced (Text) | <b>Before:</b> 73<br><b>After:</b> 41                                                                                                                                    |
| #420 (p.23 → p.26) | Replaced (Text) | <b>Before:</b> 70.6<br><b>After:</b> 37.5                                                                                                                                |
| #421 (p.23 → p.26) | Replaced (Text) | <b>Before:</b> 68.5<br><b>After:</b> 37.2                                                                                                                                |
| #422 (p.23 → p.26) | Replaced (Text) | <b>Before:</b> 67.6<br><b>After:</b> 34.4                                                                                                                                |

|                    |                 |                                                                                                                                                               |
|--------------------|-----------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------|
| #423 (p.23 → p.26) | Replaced (Text) | <b>Before:</b> 58.4<br><b>After:</b> 34.0                                                                                                                     |
| #424 (p.23 → p.26) | Replaced (Text) | <b>Before:</b> 22<br><b>After:</b> 29.7%                                                                                                                      |
| #425 (p.26)        | Added (Text)    | 25                                                                                                                                                            |
| #426 (p.24 → p.27) | Replaced (Text) | <b>Before:</b> We<br><b>After:</b> Core to SpaceXAI's mission of truth-seeking is factual correctness, informational recency, and verifiability. We therefore |
| #427 (p.27)        | Added (Text)    | in                                                                                                                                                            |
| #428 (p.27)        | Added (Text)    | is an internal benchmark that                                                                                                                                 |
| #429 (p.24 → p.27) | Replaced (Text) | <b>Before:</b> .<br><b>After:</b> . We evaluate Grok 4.6 with the Grok Build harness. Grok 4.6 scores 1.7% hallucination rate at high thinking effort.        |
| #430 (p.27)        | Added (Text)    | 26                                                                                                                                                            |
| #431 (p.28)        | Added (Text)    | GROK 4.6 MODEL CARD                                                                                                                                           |
| #432 (p.24 → p.28) | Replaced (Text) | <b>Before:</b> 18<br><b>After:</b> 22                                                                                                                         |
| #433 (p.28)        | Added (Text)    | We evaluate Grok 4.6 with the Grok Build harness. Grok 4.6 scores 81.6% at high thinking effort.                                                              |
| #434 (p.28)        | Added (Text)    | GPT-5.5 (xhigh)                                                                                                                                               |
| #435 (p.28)        | Added (Text)    | Grok 4.5 (high)                                                                                                                                               |
| #436 (p.24)        | Removed (Text)  | max)                                                                                                                                                          |
| #437 (p.24)        | Removed (Text)  | GPT-5.5 (x                                                                                                                                                    |
| #438 (p.24 → p.28) | Replaced (Text) | <b>Before:</b> Grok 4.5 (high)<br><b>After:</b> 87.8%                                                                                                         |
| #439 (p.28)        | Added (Text)    | 85.3%                                                                                                                                                         |
| #440 (p.24)        | Removed (Text)  | 63.7%                                                                                                                                                         |
| #441 (p.24)        | Removed (Text)  | 38.4%                                                                                                                                                         |
| #442 (p.24 → p.28) | Replaced (Text) | <b>Before:</b> 23<br><b>After:</b> 27                                                                                                                         |

#443 (p.25 → p.29) **Replaced** (Text)

**Before:** 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

**After:** 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 4.6 shows some cybersecurity-relevant capability gains over Grok 4.5, concentrated in cyber-defense and vulnerability-mitigation tasks rather than end-to-end offensive 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 §7.4). 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 4.6 to

#444 (p.25 → p.29) **Replaced** (Text)

**Before:** validate  
**After:** , who corroborated

#445 (p.25) **Removed** (Text)

above

#446 (p.25 → p.29) **Replaced** (Text)

**Before:** ; Cyber  
**After:** of our internal

#447 (p.25 → p.29) **Replaced** (Text)

**Before:** measure offensive and defensive security-relevant  
**After:** and testing on cyber

#448 (p.25) **Removed** (Text)

, plus calibration of cyber safety controls

#449 (p.25 → p.29) **Replaced** (Text)

**Before:** 19  
**After:** 23

#450 (p.29) **Added** (Text)

cyber

#451 (p.25 → p.29) **Replaced** (Text)

**Before:** (i.e. unmitigated by  
**After:** setting (without

#452 (p.25) **Removed** (Text)

setting

#453 (p.25 → p.29) **Replaced** (Text)

**Before:** HackerBench  
**After:** §7.4

#454 (p.29) **Added** (Text)

unsafeguarded

#455 (p.25 → p.29) **Replaced** (Text)

**Before:** 0,21,22,23 , using the unsafeguarded figures  
**After:** 4,25,26,27

|                    |                 |                                                                                                                                                                                                                                                             |
|--------------------|-----------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| #456 (p.29)        | Added (Text)    | We evaluate Grok 4.6 with the Grok Build harness. Grok 4.6 scores 79.7% at high thinking effort.                                                                                                                                                            |
| #457 (p.25 → p.29) | Replaced (Text) | <b>Before:</b> 24<br><b>After:</b> 28                                                                                                                                                                                                                       |
| #458 (p.26 → p.30) | Replaced (Text) | <b>Before:</b> 24<br><b>After:</b> 28                                                                                                                                                                                                                       |
| #459 (p.30)        | Added (Text)    | We evaluate Grok 4.6 with the Grok Build harness. Grok 4.6 scores 39.8% at high thinking effort.                                                                                                                                                            |
| #460 (p.26 → p.30) | Replaced (Text) | <b>Before:</b> under a fixed review<br><b>After:</b> . We evaluate Grok 4.6 with the Grok Build                                                                                                                                                             |
| #461 (p.30)        | Added (Text)    | Grok 4.6 scores 58.7% at high thinking effort.                                                                                                                                                                                                              |
| #462 (p.26 → p.30) | Replaced (Text) | <b>Before:</b> 25<br><b>After:</b> 29                                                                                                                                                                                                                       |
| #463 (p.27 → p.31) | Replaced (Text) | <b>Before:</b> R<br><b>After:</b> We use HackerBench, among other evaluations, metrics, and qualitative tests, to measure the effectiveness and calibration of our safeguard stack. HackerBench r                                                           |
| #464 (p.27 → p.31) | Replaced (Text) | <b>Before:</b> Harmful/Dual-Use Compliance Benign Refusal<br><b>After:</b> Grok 4.6 complies with 6.9% of dual-use requests at high thinking effort under our standard safeguard stack. Grok 4.6 has a benign refusal rate of 0.0% at high thinking effort. |
| #465 (p.27 → p.31) | Replaced (Text) | <b>Before:</b> 16.7%<br><b>After:</b> Harmful/Dual-Use Compliance                                                                                                                                                                                           |
| #466 (p.27 → p.31) | Replaced (Text) | <b>Before:</b> 0.8%<br><b>After:</b> 6.9%                                                                                                                                                                                                                   |
| #467 (p.27 → p.31) | Replaced (Text) | <b>Before:</b> 7.8%<br><b>After:</b> 7.8%                                                                                                                                                                                                                   |
| #468 (p.27)        | Removed (Text)  | 0.0%                                                                                                                                                                                                                                                        |
| #469 (p.27)        | Removed (Text)  | 1.1%                                                                                                                                                                                                                                                        |
| #470 (p.27)        | Removed (Text)  | 0.0%                                                                                                                                                                                                                                                        |
| #471 (p.31)        | Added (Text)    | Benign Refusal                                                                                                                                                                                                                                              |
| #472 (p.31)        | Added (Text)    | 0.0% 0.0%                                                                                                                                                                                                                                                   |
| #473 (p.31)        | Added (Text)    | 1.1%                                                                                                                                                                                                                                                        |
| #474 (p.31)        | Added (Text)    | 0.0%                                                                                                                                                                                                                                                        |
| #475 (p.27 → p.31) | Replaced (Text) | <b>Before:</b> 26<br><b>After:</b> 30                                                                                                                                                                                                                       |
| #476 (p.28 → p.32) | Replaced (Text) | <b>Before:</b> *<br><b>After:</b> 29                                                                                                                                                                                                                        |

|                    |                 |                                                                                                                                                                                                                                                                                                                                                                                                                |
|--------------------|-----------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| #477 (p.28 → p.32) | Replaced (Text) | <b>Before:</b> It also demonstrates no appreciable lift in dual-use capabilities compared to Grok 4.5. Weapons-path refusals appear<br><b>After:</b> Dual-use capability lift versus Grok 4.5 is noted in the biological domain but is limited (for example, a small VCT uptick); weapons-path refusals are consistent with our safeguard stack and show improvement over Grok 4.5; these results are reported |
| #478 (p.32)        | Added (Text)    | On general biological knowledge not useful for dual-use capability, Grok 4.6 shows improvement over Grok 4.5.                                                                                                                                                                                                                                                                                                  |
| #479 (p.28 → p.32) | Replaced (Text) | <b>Before:</b> any<br><b>After:</b> production                                                                                                                                                                                                                                                                                                                                                                 |
| #480 (p.28 → p.32) | Replaced (Text) | <b>Before:</b> 25<br><b>After:</b> 30                                                                                                                                                                                                                                                                                                                                                                          |
| #481 (p.28)        | Removed (Text)  | ,                                                                                                                                                                                                                                                                                                                                                                                                              |
| #482 (p.28 → p.32) | Replaced (Text) | <b>Before:</b> , so<br><b>After:</b> :                                                                                                                                                                                                                                                                                                                                                                         |
| #483 (p.28 → p.32) | Replaced (Text) | <b>Before:</b> more<br><b>After:</b> stronger performance on task distributions that overlap with                                                                                                                                                                                                                                                                                                              |
| #484 (p.28 → p.32) | Replaced (Text) | <b>Before:</b> rather than a<br><b>After:</b> , not                                                                                                                                                                                                                                                                                                                                                            |
| #485 (p.28 → p.32) | Replaced (Text) | <b>Before:</b> outcome<br><b>After:</b> refusal behavior under safeguards                                                                                                                                                                                                                                                                                                                                      |
| #486 (p.32)        | Added (Text)    | 31                                                                                                                                                                                                                                                                                                                                                                                                             |
| #487 (p.33)        | Added (Text)    | GROK 4.6 MODEL CARD                                                                                                                                                                                                                                                                                                                                                                                            |
| #488 (p.28 → p.33) | Replaced (Text) | <b>Before:</b> 25<br><b>After:</b> 30                                                                                                                                                                                                                                                                                                                                                                          |
| #489 (p.28 → p.33) | Replaced (Text) | <b>Before:</b> — 47.8%<br><b>After:</b> 44.1% 47.8%                                                                                                                                                                                                                                                                                                                                                            |
| #490 (p.28)        | Removed (Text)  | * xAI Frontier Artificial Intelligence Framework (June 30, 2026):<br><a href="https://media.x.ai/v1/website/xai-frontier-artificial-intelligence-framework-k-30-june-2026-99c40684.pdf">https://media.x.ai/v1/website/xai-frontier-artificial-intelligence-framework-k-30-june-2026-99c40684.pdf</a> .                                                                                                         |
| #491 (p.28)        | Removed (Text)  | 27                                                                                                                                                                                                                                                                                                                                                                                                             |
| #492 (p.29)        | Removed (Text)  | GROK 4.6 MODEL CARD                                                                                                                                                                                                                                                                                                                                                                                            |
| #493 (p.33)        | Added (Text)    | Dual-use refusal:                                                                                                                                                                                                                                                                                                                                                                                              |
| #494 (p.29 → p.33) | Replaced (Text) | <b>Before:</b> most borderline<br><b>After:</b> highest-severity (severity-5)                                                                                                                                                                                                                                                                                                                                  |
| #495 (p.29 → p.33) | Replaced (Text) | <b>Before:</b> split.<br><b>After:</b> subset. This consists of queries that are non-malicious but could be used in biological uplift work.                                                                                                                                                                                                                                                                    |
| #496 (p.29 → p.33) | Replaced (Text) | <b>Before:</b> Severity-5<br><b>After:</b> Dual-use                                                                                                                                                                                                                                                                                                                                                            |
| #497 (p.29 → p.33) | Replaced (Text) | <b>Before:</b> rate —<br><b>After:</b> 83.3%                                                                                                                                                                                                                                                                                                                                                                   |

|                    |                 |                                                                                                                                                                                                                                                                                                                         |
|--------------------|-----------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| #498 (p.29 → p.33) | Replaced (Text) | <b>Before:</b> 26<br><b>After:</b> 31                                                                                                                                                                                                                                                                                   |
| #499 (p.29 → p.33) | Replaced (Text) | <b>Before:</b> —<br><b>After:</b> 83.2%                                                                                                                                                                                                                                                                                 |
| #500 (p.29 → p.33) | Replaced (Text) | <b>Before:</b> 28<br><b>After:</b> 32                                                                                                                                                                                                                                                                                   |
| #501 (p.30 → p.34) | Replaced (Text) | <b>Before:</b> 27<br><b>After:</b> 32                                                                                                                                                                                                                                                                                   |
| #502 (p.30)        | Removed (Text)  | 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).        |
| #503 (p.30 → p.34) | Replaced (Text) | <b>Before:</b> 27<br><b>After:</b> 32                                                                                                                                                                                                                                                                                   |
| #504 (p.34)        | Added (Text)    | 8.7 BixBench zero-shot MCQ 33 BixBench evaluates computational-biology reasoning with analysis tools available by default, scored zero-shot. Metric Grok 4.5 (high) Grok 4.6 (high) Accuracy 93.8% 93.8% BixBench items contain no weapons-enabling content; the suite runs under Grok Build's default tool-on setting. |
| #505 (p.34)        | Added (Text)    | 33                                                                                                                                                                                                                                                                                                                      |
| #506 (p.35)        | Added (Text)    | GROK 4.6 MODEL CARD                                                                                                                                                                                                                                                                                                     |
| #507 (p.30 → p.35) | Replaced (Text) | <b>Before:</b> the refusal<br><b>After:</b> Grok 4.6's production                                                                                                                                                                                                                                                       |
| #508 (p.30 → p.35) | Replaced (Text) | <b>Before:</b> We treat r<br><b>After:</b> R                                                                                                                                                                                                                                                                            |
| #509 (p.35)        | Added (Text)    | are treated                                                                                                                                                                                                                                                                                                             |
| #510 (p.35)        | Added (Text)    | 34,35                                                                                                                                                                                                                                                                                                                   |
| #511 (p.30 → p.35) | Replaced (Text) | <b>Before:</b> -<br><b>After:</b>                                                                                                                                                                                                                                                                                       |
| #512 (p.30)        | Removed (Text)  | 28,29                                                                                                                                                                                                                                                                                                                   |
| #513 (p.30 → p.35) | Replaced (Text) | <b>Before:</b> -<br><b>After:</b>                                                                                                                                                                                                                                                                                       |
| #514 (p.35)        | Added (Text)    | -                                                                                                                                                                                                                                                                                                                       |
| #515 (p.35)        | Added (Text)    | -                                                                                                                                                                                                                                                                                                                       |
| #516 (p.30 → p.35) | Replaced (Text) | <b>Before:</b><br><b>After:</b> -                                                                                                                                                                                                                                                                                       |
| #517 (p.30 → p.35) | Replaced (Text) | <b>Before:</b> .<br><b>After:</b> attacks.                                                                                                                                                                                                                                                                              |
| #518 (p.30)        | Removed (Text)  | 29                                                                                                                                                                                                                                                                                                                      |
| #519 (p.31)        | Removed (Text)  | GROK 4.6 MODEL CARD                                                                                                                                                                                                                                                                                                     |

|                    |                 |                                                                                                                                                               |
|--------------------|-----------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------|
| #520 (p.31 → p.35) | Replaced (Text) | <b>Before:</b> —<br><b>After:</b> 1.5%                                                                                                                        |
| #521 (p.35)        | Added (Text)    | 34                                                                                                                                                            |
| #522 (p.36)        | Added (Text)    | GROK 4.6 MODEL CARD                                                                                                                                           |
| #523 (p.31 → p.36) | Replaced (Text) | <b>Before:</b> perfect<br><b>After:</b> improved                                                                                                              |
| #524 (p.31)        | Removed (Text)  | autointent                                                                                                                                                    |
| #525 (p.31 → p.36) | Replaced (Text) | <b>Before:</b> any<br><b>After:</b> one                                                                                                                       |
| #526 (p.36)        | Added (Text)    | or classifier                                                                                                                                                 |
| #527 (p.36)        | Added (Text)    | further                                                                                                                                                       |
| #528 (p.31 → p.36) | Replaced (Text) | <b>Before:</b> Depending on<br><b>After:</b> On some                                                                                                          |
| #529 (p.36)        | Added (Text)    | s                                                                                                                                                             |
| #530 (p.31 → p.36) | Replaced (Text) | <b>Before:</b> input safety controls<br><b>After:</b> safeguards                                                                                              |
| #531 (p.36)        | Added (Text)    | 0                                                                                                                                                             |
| #532 (p.31 → p.36) | Replaced (Text) | <b>Before:</b> 30<br><b>After:</b> 35                                                                                                                         |
| #533 (p.32 → p.37) | Replaced (Text) | <b>Before:</b> 4<br><b>After:</b> 3                                                                                                                           |
| #534 (p.37)        | Added (Text)    | 36                                                                                                                                                            |
| #535 (p.37)        | Added (Text)    | ; FORTRESS-RN covers radiological and nuclear items                                                                                                           |
| #536 (p.32 → p.37) | Replaced (Text) | <b>Before:</b> accuracy<br><b>After:</b> recall                                                                                                               |
| #537 (p.37)        | Added (Text)    | recall 96.7% 100.0% FORTRESS-RN R/N refusal                                                                                                                   |
| #538 (p.32 → p.37) | Replaced (Text) | <b>Before:</b> 96.7% 100.0<br><b>After:</b> * 97.9% 97.9                                                                                                      |
| #539 (p.32 → p.37) | Replaced (Text) | <b>Before:</b> 31<br><b>After:</b> * FORTRESS-RN results are taken from evaluations conducted by Scale AI.                                                    |
| #540 (p.37)        | Added (Text)    | 36                                                                                                                                                            |
| #541 (p.38)        | Added (Text)    | We test our models to ensure they refuse assistance for self-harm while still providing appropriate support and resources for users in crisis.                |
| #542 (p.33 → p.38) | Replaced (Text) | <b>Before:</b> A<br><b>After:</b> In this evaluation, a                                                                                                       |
| #543 (p.33 → p.38) | Replaced (Text) | <b>Before:</b> or offers assistance. L<br><b>After:</b> , or if it is unable to understand the intent of a user message that implies self-harm or crisis. A I |

|                    |                 |                                                                                                                                                                                                                                 |
|--------------------|-----------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| #544 (p.38)        | Added (Text)    | score                                                                                                                                                                                                                           |
| #545 (p.38)        | Added (Text)    | in this benchmark                                                                                                                                                                                                               |
| #546 (p.33 → p.38) | Replaced (Text) | <b>Before:</b> % 3.7<br><b>After:</b> 0% 0.84                                                                                                                                                                                   |
| #547 (p.38)        | Added (Text)    | for instance,                                                                                                                                                                                                                   |
| #548 (p.33 → p.38) | Replaced (Text) | <b>Before:</b> 30<br><b>After:</b> 37                                                                                                                                                                                           |
| #549 (p.33 → p.38) | Replaced (Text) | <b>Before:</b> L<br><b>After:</b> A I                                                                                                                                                                                           |
| #550 (p.38)        | Added (Text)    | score                                                                                                                                                                                                                           |
| #551 (p.38)        | Added (Text)    | in this benchmark                                                                                                                                                                                                               |
| #552 (p.33 → p.38) | Replaced (Text) | <b>Before:</b> 3.8<br><b>After:</b> 1.90                                                                                                                                                                                        |
| #553 (p.33 → p.38) | Replaced (Text) | <b>Before:</b> 32<br><b>After:</b> 37                                                                                                                                                                                           |
| #554 (p.34 → p.39) | Replaced (Text) | <b>Before:</b> 33<br><b>After:</b> 38                                                                                                                                                                                           |
| #555 (p.40)        | Added (Text)    | GROK 4.6 MODEL CARD                                                                                                                                                                                                             |
| #556 (p.40)        | Added (Text)    | Acknowledgements We thank our evaluation partners for running benchmarks, sharing results, and helping us measure Grok 4.6 against real-world workloads: • Abundant AI • Artificial Analysis • Atopile • Cognition • Databricks |
| #557 (p.40)        | Added (Text)    | • Datacurve • gNucleus • Harbor • Harvey • Mecado                                                                                                                                                                               |
| #558 (p.40)        | Added (Text)    | • Mercor • Normal Factory • Scale AI • Vals AI                                                                                                                                                                                  |
| #559 (p.40)        | Added (Text)    | 39                                                                                                                                                                                                                              |
| #560 (p.41)        | Added (Text)    | SpaceXAI Acceptable Use Policy.<br><a href="https://x.ai/legal/acceptable-use-policy">https://x.ai/legal/acceptable-use-policy</a> 2.                                                                                           |
| #561 (p.35)        | Removed (Text)  | AnySphere /                                                                                                                                                                                                                     |
| #562 (p.35)        | Removed (Text)  |                                                                                                                                                                                                                                 |
| #563 (p.41)        | Added (Text)    |                                                                                                                                                                                                                                 |
| #564 (p.35 → p.41) | Replaced (Text) | <b>Before:</b> 2<br><b>After:</b> 3                                                                                                                                                                                             |
| #565 (p.35 → p.41) | Replaced (Text) | <b>Before:</b> 3<br><b>After:</b> 4. Terminus-2. Harbor Framework.<br><a href="https://www.harborframework.com/docs/agents/terminus-2">https://www.harborframework.com/docs/agents/terminus-2</a> 5                             |
| #566 (p.35 → p.41) | Replaced (Text) | <b>Before:</b> 4<br><b>After:</b> 6                                                                                                                                                                                             |
| #567 (p.35 → p.41) | Replaced (Text) | <b>Before:</b> 5<br><b>After:</b> 7. mini-swe-agent. SWE-agent.<br><a href="https://github.com/swe-agent/mini-swe-agent">https://github.com/swe-agent/mini-swe-agent</a> 8                                                      |

|                                    |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                 |
|------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| #568 (p.35 → p.41) Replaced (Text) | <b>Before:</b> 6<br><b>After:</b> 9                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                             |
| #569 (p.35 → p.41) Replaced (Text) | <b>Before:</b> 7. EEBench. atopile, 2026. <a href="https://eebench.org/">https://eebench.org/</a> 8. 3DCodeBench. Gao, Shu, Ye, Xiong, Makadia, Guo, Itti, and Chen, 2026.<br><b>After:</b> ·                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                   |
| #570 (p.35 → p.41) Replaced (Text) | <b>Before:</b> 2606.01057 9. CadGenBench. Mecado / Hugging Face, 2026. <a href="https://github.com/huggingface/cadgenbench">https://github.com/huggingface/cadgenbench</a> 10. CadBench (Parametric CAD Bench). gNucleus. <a href="https://www.gnucleus.ai/cad-bench">https://www.gnucleus.ai/cad-bench</a> 11.<br><b>After:</b> 2601.11868 10.                                                                                                                                                                                                                                                                                                                                                                                                                                                 |
| #571 (p.41) Added (Text)           | Patwardhan et al., 2025.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                        |
| #572 (p.35 → p.41) Replaced (Text) | <b>Before:</b> 12<br><b>After:</b> 11                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                           |
| #573 (p.35 → p.41) Replaced (Text) | <b>Before:</b><br><b>After:</b> -                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                               |
| #574 (p.35 → p.41) Replaced (Text) | <b>Before:</b> 13<br><b>After:</b> 12                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                           |
| #575 (p.35 → p.41) Replaced (Text) | <b>Before:</b> 14. Vals Index<br><b>After:</b> 13. OfficeQA Pro. Opsahl-Ong et al. (Databricks), 2026. <a href="https://arxiv.org/abs/2603.08655">https://arxiv.org/abs/2603.08655</a> · <a href="https://github.com/databricks/officeqa">https://github.com/databricks/officeqa</a> 14. Legal Agent Benchmark (LAB / HLab). Harvey, 2026                                                                                                                                                                                                                                                                                                                                                                                                                                                       |
| #576 (p.35 → p.41) Replaced (Text) | <b>Before:</b> ·<br><b>After:</b> third-party leaderboard:                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                      |
| #577 (p.35 → p.41) Replaced (Text) | <b>Before:</b> vals_index · <a href="https://www.vals.ai/benchmarks">https://www.vals.ai/benchmarks</a> 15. OfficeQA Pro. Databricks<br><b>After:</b> hlab · Harvey announcement: <a href="https://www.harvey.ai/blog/introducing-harveys-legal-agent-benchmark">https://www.harvey.ai/blog/introducing-harveys-legal-agent-benchmark</a> 15. Valkyrie. Vals AI agentic evaluation framework. <a href="https://github.com/vals-ai/Valkyrie">https://github.com/vals-ai/Valkyrie</a> · HLab on Valkyrie: <a href="https://www.vals.ai/benchmarks/hlab">https://www.vals.ai/benchmarks/hlab</a> 16. EEBench. atopile, 2026. <a href="https://eebench.org/">https://eebench.org/</a> 17. 3DCodeBench. Gao, Shu, Ye, Xiong, Makadia, Guo, Itti, and Chen                                            |
| #578 (p.35 → p.41) Replaced (Text) | <b>Before:</b> 2603.08655 · <a href="https://github.com/databricks/officeqa">https://github.com/databricks/officeqa</a> 16. Harvey Legal Agent Benchmark (LAB), Vals implementation. Vals AI, 2026. <a href="https://www.vals.ai/benchmarks/hlab">https://www.vals.ai/benchmarks/hlab</a> 17<br><b>After:</b> 2606.01057 18. PartBench. Normal Factory, 2026. <a href="https://www.normalfactory.com/">https://www.normalfactory.com/</a> 19. CADGenBench. Mecado / Hugging Face, 2026. <a href="https://www.mecado.com/">https://www.mecado.com/</a> · <a href="https://github.com/huggingface/cadgenbench">https://github.com/huggingface/cadgenbench</a> 20. CADBench (Parametric CAD Bench). gNucleus. <a href="https://www.gnucleus.ai/cad-bench">https://www.gnucleus.ai/cad-bench</a> 21 |
| #579 (p.35 → p.41) Replaced (Text) | <b>Before:</b> 18<br><b>After:</b> 22                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                           |
| #580 (p.35 → p.41) Replaced (Text) | <b>Before:</b> 19<br><b>After:</b> 23                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                           |
| #581 (p.35 → p.41) Replaced (Text) | <b>Before:</b> 20<br><b>After:</b> 40                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                           |

|                    |                 |                                                                                                                                                                                                                                                                                                                                                  |
|--------------------|-----------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| #582 (p.42)        | Added (Text)    | GROK 4.6 MODEL CARD                                                                                                                                                                                                                                                                                                                              |
| #583 (p.42)        | Added (Text)    | 24                                                                                                                                                                                                                                                                                                                                               |
| #584 (p.35 → p.42) | Replaced (Text) | <b>Before:</b> 21<br><b>After:</b> 25                                                                                                                                                                                                                                                                                                            |
| #585 (p.35 → p.42) | Replaced (Text) | <b>Before:</b> 22<br><b>After:</b> 26                                                                                                                                                                                                                                                                                                            |
| #586 (p.35 → p.42) | Replaced (Text) | <b>Before:</b> 34<br><b>After:</b> 27                                                                                                                                                                                                                                                                                                            |
| #587 (p.36)        | Removed (Text)  | GROK 4.6 MODEL CARD                                                                                                                                                                                                                                                                                                                              |
| #588 (p.36)        | Removed (Text)  | 23                                                                                                                                                                                                                                                                                                                                               |
| #589 (p.36 → p.42) | Replaced (Text) | <b>Before:</b> 24<br><b>After:</b> 28                                                                                                                                                                                                                                                                                                            |
| #590 (p.36 → p.42) | Replaced (Text) | <b>Before:</b> 25<br><b>After:</b> 29. xAI Frontier Artificial Intelligence Framework (FAIF). xAI, June 30, 2026. <a href="https://media.x.ai/v1/website/xai-frontier-artificial-intelligence-framework-30-june-2026-99c40684.pdf">https://media.x.ai/v1/website/xai-frontier-artificial-intelligence-framework-30-june-2026-99c40684.pdf</a> 30 |
| #591 (p.36 → p.42) | Replaced (Text) | <b>Before:</b> 26<br><b>After:</b> 31                                                                                                                                                                                                                                                                                                            |
| #592 (p.36 → p.42) | Replaced (Text) | <b>Before:</b> 27<br><b>After:</b> 32                                                                                                                                                                                                                                                                                                            |
| #593 (p.36 → p.42) | Replaced (Text) | <b>Before:</b> 28<br><b>After:</b> 33. BixBench. Mitchener, Laurent, et al., 2025. <a href="https://github.com/Future-House/BixBench">https://github.com/Future-House/BixBench</a> · <a href="https://arxiv.org/abs/2503.00096">https://arxiv.org/abs/2503.00096</a> 34                                                                          |
| #594 (p.36 → p.42) | Replaced (Text) | <b>Before:</b> 29<br><b>After:</b> 35                                                                                                                                                                                                                                                                                                            |
| #595 (p.36 → p.42) | Replaced (Text) | <b>Before:</b> 30<br><b>After:</b> 36. FORTRESS. Knight, Deshpande, et al. (Scale AI), 2025. <a href="https://labs.scale.com/leaderboard/fortress">https://labs.scale.com/leaderboard/fortress</a> · <a href="https://arxiv.org/abs/2506.14922">https://arxiv.org/abs/2506.14922</a> 37                                                          |
| #596 (p.36 → p.42) | Replaced (Text) | <b>Before:</b> 35<br><b>After:</b> 41                                                                                                                                                                                                                                                                                                            |