HARVARDNETWORK / INDEPENDENT RESEARCH

Grok 4.6: Independent Evaluation White Paper

Frontier reasoning at a mid-tier price, with half the context of most direct rivals. Include in evaluations where top-tier reasoning is needed but Opus pricing is difficult to justify. Add a context-overflow test because the 500k limit is the cohort's smallest frontier window. This conclusion is a procurement hypothesis, not a universal ranking: the report synthesizes public evidence and does not claim to have run a private laboratory evaluation.

Frontier Model Evidence Review · Edition 2026.08 · Evidence cutoff 28 August 2026

13 CHAPTERS / 7 EVIDENCE TABLES

This is a structured synthesis of public evidence, not a claim of original laboratory testing.

Executive Summary

***Evaluation subject**: Grok 4.6 (generally available; exact identifier: grok-4.6)*
***Benchmark cohort**: GPT-5.6 Sol, Claude Opus 5, Gemini 3.7 Flash, DeepSeek-V4-Pro, Llama 4 Maverick, Qwen3.8-Max, Kimi K3, GLM-5.3, MiniMax M3*
***Evidence cutoff**: 2026-08-28*

Bottom line

Frontier reasoning at a mid-tier price, with half the context of most direct rivals. Include in evaluations where top-tier reasoning is needed but Opus pricing is difficult to justify. Add a context-overflow test because the 500k limit is the cohort's smallest frontier window. This conclusion is a procurement hypothesis, not a universal ranking: the report synthesizes public evidence and does not claim to have run a private laboratory evaluation.

At the 28 August 2026 cutoff, Grok 4.6 scored **61** on Artificial Analysis Intelligence Index v4.1.1 at **high reasoning effort**, ranking **3 of 10** in this selected cohort when ties are ordered by measured output speed. The same operator reported approximately **57.0 output tokens per second**, a **0.50M-token context window**, and proprietary availability [S01] [S02] [S12]. The index operator estimates the composite's 95% confidence interval at less than ±1 point from repeated-model experiments, while warning that individual evaluation intervals may be wider [S01]. We therefore treat one-point gaps as directional rather than decisive.

Decision signal	Evidence at cutoff
Composite capability	61 / 100 on AA Index v4.1.1
Output throughput	57.0 tokens/s
Context	0.50M tokens
Input modalities in common measurement	text and image
Access model	proprietary
Public price basis	$2 input / $6 output per 1M tokens

What the evidence supports

The strongest case is frontier reasoning under a moderate token budget, tool-using agents, and image-grounded analysis. The principal advantages are Tied second-highest common-harness score, Mid-tier price, and Strong measured terminal and agentic results. Artificial Analysis measured a score of 61 at high effort and about 57 output tokens per second, with strong cited Terminal-Bench and banking-agent results. These component results are still conditioned on a particular harness and effort level. Vendor materials establish identity, availability, architecture disclosures and price, but vendor benchmark tables are not used as the primary comparative result [S08] [S12].

What it does not support

The public record does not establish universal reliability, truthfulness, production safety or return on investment. The major deployment risks are Smaller context than most peers, Opaque internals, and New release with limited longitudinal evidence. A composite English-language benchmark cannot substitute for tests using the buyer's prompts, tools, languages, permissions, latency targets and review process.

Recommended decision

Proceed only through a task-level bake-off with a frozen model identifier, reasoning setting, harness and price snapshot. Start with shadow use; require human approval for consequential actions; record failures, retries, latency and accepted-output cost. Avoid inputs requiring more than 500k context, open-weight requirements, and workflows needing mature public safety documentation. The rest of this paper explains the evidence and the limits behind that recommendation.

Research Background

Model selection

This review covers **Grok 4.6** (`grok-4.6`), released in August 2026 and generally available. It was selected because Grok 4.6 is the latest released general-purpose Grok model at the evidence cutoff. The cutoff rule matters: announced models that were not publicly usable by 28 August 2026 were excluded. The identity and release status come from the vendor's own publication, where first-party evidence is the appropriate source [S08].

xAI does not disclose parameter count or training mix. Grok 4.6 exposes reasoning-effort controls, image input and a 500,000-token context window. These are product and architecture disclosures, not independently audited training facts. Where the vendor withholds parameters, data mix or compute, this paper records the absence rather than estimating them.

The decision problem

Model selection is no longer a one-dimensional search for the highest benchmark number. A production system combines a model version, inference setting, provider route, prompt, tools, retrieval system, data permissions, retry policy and human review. Changing any of these can change measured quality, latency, cost and risk. Terminal-Bench makes this coupling visible by publishing the agent harness, model, reasoning effort, accuracy, uncertainty and run cost for each submission [S03]. SWE-bench similarly distinguishes benchmark variants and whether a run was directly checked by its team [S04].

For Grok 4.6, the procurement question is therefore: **does its particular mix of capability, throughput, context, modalities, access and price produce more accepted work on the target workflow than current alternatives?** The common evidence gives an initial screen. It does not answer that organization-specific question.

Research questions

1. How does Grok 4.6 compare with the other nine selected flagship or mainstream family representatives under one current, common harness?
2. Which deployment conditions are supported by its measured score of 61, measured throughput of 57.0 tokens per second and 0.50M context?
3. Which claims originate with the vendor, and which have independent support?
4. What costs and operational controls are missing from token-price comparisons?
5. What evidence would be required before high-impact or autonomous deployment?

Evidence context

The model market changes faster than normal publication cycles. This report is a dated evidence snapshot, not a permanent league table. Its comparison set intentionally mixes proprietary and open-weight systems because buyers face both choices; it does not imply that hosted API price and self-hosted total cost are equivalent. It also keeps current Index v4.1.1 results separate from older index versions, a particularly important control for long-lived releases.

Practical Recommendations

Recommended posture

Include in evaluations where top-tier reasoning is needed but Opus pricing is difficult to justify. Add a context-overflow test because the 500k limit is the cohort's smallest frontier window. The evidence supports a **qualified shortlist**, not an unconditional deployment. Grok 4.6 should earn production traffic by outperforming a cheaper or simpler baseline on the buyer's accepted-output metric.

- **Use:** Frontier reasoning under a moderate token budget.
- **Use:** Tool-using agents.
- **Use:** Image-grounded analysis.
- **Do not default to it for:** Inputs requiring more than 500k context.
- **Do not default to it for:** Open-weight requirements.
- **Do not default to it for:** Workflows needing mature public safety documentation.

Configure the evaluation before selecting the model

Freeze the exact model identifier (`grok-4.6`), high reasoning effort, maximum output, tool permissions and retry limit. Record end-to-end latency rather than output speed alone. The reported 57.0 tokens per second excludes queueing, prompt processing, tool calls and human review [S12]. For long-context tests, use realistic retrieval noise; a 0.50M limit proves capacity, not reliable use of every token.

Three-stage rollout

1. **Offline acceptance test.** Sample at least 100 representative tasks, stratified by difficulty and risk. Blind-review outputs against a current production baseline. Record pass/fail, severity, latency, input and output tokens, retries and reviewer minutes.
2. **Shadow production.** Send live inputs to the candidate without allowing it to act. Compare drift, refusal behavior, tool-call plans and cost. Red-team prompt injection and data-exfiltration paths.
3. **Bounded activation.** Grant the minimum permissions needed, require approval for high-impact steps and define automatic rollback thresholds. Audit both successful and failed trajectories.

Decision rule

Select the model only when the confidence interval around the **local task acceptance rate** clears a predeclared practical threshold, not merely when it wins by a few public benchmark points. For a binary local metric, report a Wilson interval and the exact sample size. Public Index v4.1.1 should be treated as a prior for shortlist formation, not as the acceptance test itself [S01].

Evaluation Methodology

Study design

This is a structured secondary-research evaluation. No new model inference was conducted for publication. The workflow followed four stages: citation-standard review, source outlining, source-to-chapter mapping and chapter drafting. Each report uses twelve source cards: seven common methodology, benchmark and governance sources plus five model-specific official or independent sources. The full cards and mapping are retained in the production directory.

Evidence hierarchy

- **T1 - primary methodology and public standards:** benchmark papers, operator documentation and government frameworks.
- **T2 - independent measurement:** results produced by a benchmark operator under a disclosed common harness.
- **T3 - vendor documentation:** authoritative for model identity, product limits, architecture disclosures, availability and list price; comparative performance claims remain vendor claims.
- **T4-T6 - secondary reporting, community evidence and commentary:** useful for leads and failure hypotheses, but not used here as the primary quantitative result.

The report gives precedence to T1 and T2 evidence for comparative claims. Vendor sources are necessary for facts only the provider controls, while their benchmark claims are labelled and not substituted for independent measurement. This is consistent with the adopted accuracy rule: material claims require source authority, context and clear uncertainty.

Common comparison

The quantitative anchor is Artificial Analysis Intelligence Index **v4.1.1**, observed on **28 August 2026**. It combines nine evaluations across agents (34%), coding (24%), scientific reasoning (24%) and general capability (18%). The operator reports standardized prompting, model-appropriate reasoning settings, pass@1 scoring, repeated trials on several component evaluations and a stated composite 95% confidence interval of less than ±1 point based on experiments with more than ten repeats on certain models [S01]. The paper does not reinterpret that statement as a model-specific interval.

Grok 4.6 was tested at **high reasoning effort**. That condition is part of the result. Reasoning effort changes quality, latency and cost; comparing a max-effort result with another model's low-effort result would answer a different question. The public leaderboards and model analysis were cross-checked for score, throughput, context, modalities and weight availability [S02] [S12].

Comparability controls

The cohort is restricted to one latest publicly released flagship or mainstream model from each user-requested family. Preview-only and unreleased models are excluded. All composite scores use the same current index version. We do not mix vendor benchmark tables with the common leaderboard, do not treat older index scores as current and do not infer a missing number from a chart.

Throughput is reported as output tokens per second from the independent operator. It is not end-to-end latency. Context is the supported input capacity reported or observed by sources; it is not a long-context accuracy score. Modalities indicate accepted inputs, not equal quality across text, image, audio or video.

Cost method

List prices are timestamped and reported without pretending they are total cost of ownership. The illustrative workload uses one million input tokens and 250,000 output tokens, calculated as:

```
illustrative cost = input price + 0.25 × output price
```

Cache discounts, batch discounts, regional taxes, tool fees, storage, retries, reviewer time and self-hosting infrastructure are excluded unless explicitly stated. An absent canonical price remains absent.

Safety method

Safety is evaluated as a disclosure-and-control question, not a single score. The analysis separates vendor documentation from operational evidence and maps deployment recommendations to NIST AI RMF's Govern, Map, Measure and Manage functions [S05]. The NIST Generative AI Profile supplies a cross-sector risk taxonomy [S06]. A model is not declared safe merely because a provider publishes a system card or acceptable-use policy.

Reproducibility boundary

The report records exact model ID, date, setting, index version, source URLs and formulas. It cannot reproduce the benchmark operator's private datasets, provider routing or undisclosed model internals. Readers should re-check live pages before purchase and rerun their local evaluation after a provider alias, price or system behavior changes.

Evaluation Metrics

Metric set

Metric	Definition in this report	Correct interpretation	Common misuse
Intelligence Index	Weighted v4.1.1 composite across nine evaluations	Shortlisting signal under one harness	Universal intelligence or product quality
Output tokens/s	Independent measured generation throughput	Streaming/output phase speed	Full response latency or task duration
Context tokens	Maximum supported input capacity	Upper bound for request design	Proof of accurate recall across the whole window
Token price	Dated public price per million tokens	One component of variable cost	Total cost of accepted work
Modalities	Accepted input types in the cited product or measurement	Integration surface	Equal competence in every modality
Weight access	Proprietary or downloadable weights	Degree of deployment control	Complete freedom from license or infrastructure constraints

The common index weights agentic work heavily: GDPval-AA v2 contributes 20%, τ³-Banking 14%, Terminal-Bench v2.1 16%, SciCode 8%, HLE 12%, GPQA Diamond 6%, CritPt 6%, AA-LCR 6% and AA-Omniscience 12% [S01]. That construction is useful for modern agent workloads, but organizations with different task mixes should not inherit the weights uncritically.

Grok 4.6 measurement card

- **Setting:** high reasoning effort
- **Index:** 61
- **Output throughput:** 57.0 tokens/s
- **Context:** 0.50M tokens
- **Measured input modalities:** text and image
- **Weights:** proprietary

The operator's composite uncertainty statement-less than ±1 point at 95% confidence-means one-point rank gaps should not drive procurement [S01]. It does not erase larger gaps, but practical significance still depends on the task. The report deliberately avoids false precision: no model-level confidence interval is shown because the source does not publish one for this exact row.

Local metrics to add

A production evaluation should add task acceptance rate, severe-error rate, tool-call success, citation correctness, instruction retention, time to first token, wall-clock completion, reviewer minutes and cost per accepted task. Safety-critical workflows should use scenario-specific harm metrics and severity-weighted failure counts. These measures make the public comparison actionable without claiming that one composite can represent every deployment.

Core Capability Results

Independent result first

Grok 4.6 recorded **61** on Intelligence Index v4.1.1 at **high reasoning effort**, placing it 3 of 10 in the selected cohort under the report's tie rule [S12]. Its measured output throughput was **57.0 tokens per second**. These two numbers describe different qualities: a high score cannot guarantee responsiveness, while fast generation cannot repair a wrong answer.

Artificial Analysis measured a score of 61 at high effort and about 57 output tokens per second, with strong cited Terminal-Bench and banking-agent results. These component results are still conditioned on a particular harness and effort level.

Capability profile

- **1.** Tied second-highest common-harness score.
- **2.** Mid-tier price.
- **3.** Strong measured terminal and agentic results.

The model supports a **0.50M-token context window** and the common evidence records text and image inputs. xAI does not disclose parameter count or training mix. Grok 4.6 exposes reasoning-effort controls, image input and a 500,000-token context window. The distinction between capacity and use is essential: long-context support does not show uniform retrieval, ordering or reasoning accuracy across that window. A buyer should test realistic document collections with distractors, conflicting passages and information located near the beginning, middle and end.

Agentic and coding interpretation

Index v4.1.1 assigns 58% of its weight to agentic and coding categories [S01]. That makes the result more relevant to tool-using systems than older knowledge-heavy aggregates, yet it also increases harness sensitivity. Terminal-Bench publishes model and agent separately and shows materially different outcomes for different combinations [S03]. SWE-bench distinguishes standard, verified, multilingual and multimodal tracks [S04]. A model score should never be copied into a claim about a specific coding product without the matching harness.

Scientific and knowledge interpretation

Scientific reasoning accounts for 24% of the index, including Humanity's Last Exam, GPQA Diamond and CritPt. HLE is intentionally difficult and expert-authored [S07]. General capability contributes the remaining 18% through long-context reasoning and an omniscience/hallucination measure. This breadth reduces dependence on one dataset, but the suite is still primarily text-based and English-language [S01].

Product evidence

The vendor documents Grok 4.6 as generally available under the identifier `grok-4.6` [S08]. Vendor materials are used here for availability, configuration, modality and pricing facts. Their own benchmark tables are contextual evidence, not the primary comparative result, because test prompts, internal harnesses and selection rules may differ from peer submissions.

Boundaries attached to the result

- **Boundary 1.** Smaller context than most peers.
- **Boundary 2.** Opaque internals.
- **Boundary 3.** New release with limited longitudinal evidence.

No public benchmark result in this review establishes factual reliability on current events, safe autonomous operation, performance on a private corpus or compliance with a regulated workflow. Those remain validation tasks for the deployer.

Competitor Comparison

Same-version cohort

Model	AA Index v4.1.1	Output tok/s	Context	Weights
Claude Opus 5	63	54.0	1M	proprietary
GPT-5.6 Sol	61	73.6	1.05M	proprietary
Grok 4.6	61	57.0	0.50M	proprietary
GLM-5.3	60	66.5	1M	open weights
Kimi K3	60	39.0	1.05M	open weights
Qwen3.8-Max	58	20.7	1M	open weights
Gemini 3.7 Flash	56	329.7	1M	proprietary
DeepSeek-V4-Pro	53	68.0	1M	open weights
MiniMax M3	45	117.9	1M	open weights with commercial restrictions
Llama 4 Maverick	14	100.0	1M	open weights

Source: Artificial Analysis Index v4.1.1 model pages and leaderboard, observed 28 August 2026 [S01] [S02]. Throughput is not end-to-end latency. The table is not a safety ranking.

Grok 4.6 scores below Claude Opus 5. Models with higher observed output throughput include GPT-5.6 Sol, GLM-5.3, Gemini 3.7 Flash, and DeepSeek-V4-Pro. These comparisons narrow the shortlist; they do not determine which system completes a particular workflow at lowest accepted-output cost.

Position by procurement objective

- **Maximum common-harness capability:** Claude Opus 5 leads this snapshot at 63, with GPT-5.6 Sol and Grok 4.6 at 61. One-point differences should be treated cautiously because the operator states a composite interval below ±1 point, not zero [S01].
- **Open-weight frontier:** Kimi K3 and GLM-5.3 each reach 60; Qwen3.8-Max follows at 58. Their deployment control comes with license, hardware and operations work.
- **Interactive throughput:** Gemini 3.7 Flash is the clear output-speed outlier at roughly 330 tokens per second. MiniMax M3 and Llama 4 Maverick are faster than most other open models, but capability differs sharply.
- **Low published API price:** DeepSeek-V4-Pro and MiniMax M3 publish aggressive pricing conditions. Cache policy, context tier and discounts make headline comparisons unstable.

Direct decision route for Grok 4.6

Include in evaluations where top-tier reasoning is needed but Opus pricing is difficult to justify. Add a context-overflow test because the 500k limit is the cohort's smallest frontier window. Compare at least one stronger-scoring model, one lower-cost model and-where relevant-one open-weight model. Hold prompt, tools, data and review rubric constant. Publish both task success and wall-clock/cost distributions, not only an average.

Why no single winner is declared

The cohort contains different access models, modality surfaces, context limits and reasoning settings. It also lacks a common safety and factuality suite for this exact set of releases. A ranked score table is valuable evidence, but declaring a universal winner would exceed what the sources can support.

Failure Cases and Boundaries

Evidence boundary

This review did not execute Grok 4.6 and therefore does not present invented transcripts as observed failures. Instead it converts limitations in the public record into falsifiable tests. This distinction matters: a missing evaluation is an evidence gap, not proof that the model fails; a vendor demonstration is not proof that it succeeds under production conditions.

1. Smaller context than most peers

This risk is treated as a deployment hypothesis to test, not as a fabricated incident record. Define a targeted scenario, expected safe behavior, severity level and rollback threshold before activation.

2. Opaque internals

This risk is treated as a deployment hypothesis to test, not as a fabricated incident record. Define a targeted scenario, expected safe behavior, severity level and rollback threshold before activation.

3. New release with limited longitudinal evidence

This risk is treated as a deployment hypothesis to test, not as a fabricated incident record. Define a targeted scenario, expected safe behavior, severity level and rollback threshold before activation.

4. Long-context degradation

The advertised 0.50M context is a capacity limit. Test retrieval from different positions, contradictory evidence, duplicated instructions and irrelevant bulk. Score citation precision and whether the model admits when the requested evidence is absent.

5. Agent-harness dependence

Terminal-Bench's published results attach an agent and effort setting to every model result [S03]. Recreate realistic tool failures: stale credentials, partial writes, ambiguous confirmations, rate limits and malicious content returned by tools. The model should pause, preserve state and request approval rather than improvise destructive actions.

6. Benchmark-to-workflow transfer

The common index is primarily English and text based [S01]. Test the actual languages, file types, domain vocabulary and output constraints in production. For code, distinguish issue resolution from greenfield generation and from repository maintenance; SWE-bench itself maintains separate tracks because these are not interchangeable [S04].

Required failure log

For every local run, record model ID, date, provider, effort, prompt hash, tool versions, tokens, latency, retries, result, reviewer label and failure severity. Preserve near misses and refusals as well as successes. The resulting distribution is more useful than a curated gallery of favorable examples.

Cost-Effectiveness

Price snapshot

Item	Value
Input price	$2 per 1M tokens
Output price	$6 per 1M tokens
Basis	xAI API list price at cutoff
Illustrative workload	1M input + 250k output tokens
Calculation	$2 + 0.25 × $6 = $3.50

At the dated price basis, one million input tokens plus 250,000 output tokens costs approximately **$3.50** before discounts, tools, retries and human review. The price is a 28 August 2026 snapshot, not a quote. Provider region, caching, batch mode, context tier and negotiated terms can change it [S08] [S10].

Throughput and time

The independent operator measured **57.0 output tokens per second** [S12]. A 2,000-token answer would therefore spend roughly **35.1 seconds** in the output phase under that measurement. This estimate excludes queueing, input processing, reasoning tokens that may be billed or hidden, tool calls and retries. It should not be presented as predicted application latency.

Cost per accepted task

The economically relevant measure is:

```
(model + tool + infrastructure + review + retry cost) / accepted tasks
```

A cheaper model can cost more if it requires extra attempts or reviewer repair. A premium model can be economical if it reduces severe failures or expensive review. For agents, cap maximum turns and tool spending; record failed trajectories because success-only cost systematically understates deployment expense.

Hosted-service accounting

For proprietary hosted access, include regional availability, data-retention controls, tool and storage charges, observability, retries, rate-limit headroom, review labor and provider-switching cost. The vendor bears most inference infrastructure, but the token invoice is still not the total operating cost.

Safety and Alignment

What can be concluded

Public product and safety materials show that the vendor has documented at least part of the deployment surface [S08] [S10]. They do not establish that Grok 4.6 is safe for every use. Safety depends on the model, system prompt, tools, data, permissions, user population and monitoring. For this access model, responsibility is shared by the **model provider and deployment operator**.

The NIST AI RMF frames risk work as four continuous functions: Govern, Map, Measure and Manage [S05]. Its Generative AI Profile adds cross-sector considerations for confabulation, harmful content, information integrity, privacy, security, bias and human-AI configuration [S06]. This report uses those documents as a control framework rather than claiming regulatory certification.

Minimum control profile

Function	Required control before production
Govern	Named owner, approved uses, vendor/version register, incident and change policy
Map	Data flows, affected people, threat actors, permissions, failure severity and fallback
Measure	Task acceptance, severe errors, injection resistance, privacy leakage, bias and refusal tests
Manage	Least privilege, approval gates, rate/spend limits, logging, rollback and user recourse

Model-specific priorities

The principal risk hypotheses are Smaller context than most peers, Opaque internals, and New release with limited longitudinal evidence. These should drive the red-team suite. Where weights are downloadable, the deployer gains inspection and hosting control but also assumes more responsibility for serving security, abuse prevention, updates and model modifications. Where the model is proprietary, the provider controls internals and routing while the customer must still govern prompts, retrieval, tools and downstream decisions.

High-impact deployment boundary

Do not allow the model to make final decisions in health, employment, credit, education, legal rights, critical infrastructure or physical safety solely on the evidence reviewed here. Require qualified human review, traceable source material, appeal or override paths and scenario-specific legal assessment. A general benchmark score is not a validated high-impact performance claim.

Monitoring

Re-run safety and task tests after any alias update, provider migration, prompt change, tool addition, retrieval-index change or policy update. Monitor refusal drift, unexpected tool calls, sensitive-data exposure and reviewer disagreement. Publish incident counts with denominators; raw counts without workload volume can mislead.

Dataset and Evidence Description

Quantitative suite

Intelligence Index v4.1.1 contains nine evaluations [S01]:

Category	Evaluation	Items / repeats disclosed by operator	Index weight
Agents	GDPval-AA v2	220 tasks, one run	20%
Agents	τ^3-Banking	97 tasks, five repeats	14%
Coding	Terminal-Bench v2.1	89 tasks, three repeats	16%
Coding	SciCode	288 test subproblems, three repeats	8%
Scientific	Humanity's Last Exam	2,158 items, one run	12%
Scientific	GPQA Diamond	198 items, five repeats	6%
Scientific	CritPt	70 items, five repeats	6%
General	AA-LCR	100 items, three repeats	6%
General	AA-Omniscience	6,000 items, one run	12%

The operator describes the suite as primarily English-language and text-based. Image, speech and multilingual capabilities are benchmarked separately and are not represented by the headline index [S01]. HLE's expert-authored difficult questions broaden academic coverage but do not represent ordinary enterprise task frequency [S07].

Source corpus

This report's twelve source cards consist of two independent common-harness sources, two coding-evaluation operator sources, two NIST governance sources, one benchmark paper, and five model-specific sources. Vendor sources establish product facts; independent sources anchor comparative measurements. Every card records author, year, tier, admissible use and limitation.

Model record

The exact subject is `grok-4.6` at high reasoning effort, measured on the report's 2026-08-28 snapshot. The record includes score 61, throughput 57.0, context 0.50M, modalities text and image and proprietary status. No private prompt-response dataset was created for this paper.

Research Limitations

This paper is a source-based evaluation, not an original laboratory benchmark. Its strongest quantitative evidence comes from one independent operator; methodological transparency reduces but does not remove operator dependence. Private evaluation items and provider routing cannot be fully reproduced.

The comparison is dated 28 August 2026. Model aliases, prices, safety policies and leaderboards can change. Results apply to **high reasoning effort** and should not be transferred to another effort level. The index is primarily English and text based, so it underrepresents multilingual and multimodal deployment needs [S01].

The report does not publish a model-specific confidence interval because the source provides only a suite-level estimate. It does not normalize self-hosted and API total cost, measure energy use, inspect training data, audit weights, or establish legal compliance. Context capacity is not long-context accuracy. Throughput is not end-to-end latency. Vendor safety documentation is not independent assurance.

Most importantly, public benchmarks do not reveal performance on the reader's private tasks. The recommendation is therefore conditional: Include in evaluations where top-tier reasoning is needed but Opus pricing is difficult to justify. Add a context-overflow test because the 500k limit is the cohort's smallest frontier window. A local, pre-registered evaluation remains necessary.

Appendix and References

Reproducibility record

Field	Recorded value
Report edition	2026.08
Evidence cutoff	2026-08-28
Model	Grok 4.6
Exact identifier	`grok-4.6`
Release status	generally available
Independent benchmark	Artificial Analysis Intelligence Index v4.1.1
Evaluation setting	high reasoning effort
Index result	61
Output throughput	57.0 tokens/s
Context	0.50M tokens
Access	proprietary
Price basis	xAI API list price at cutoff

Formulae

- Illustrative token cost: `1 × input price + 0.25 × output price`.
- Local acceptance rate: `accepted tasks / attempted tasks` with exact sample size and Wilson interval.
- Accepted-task cost: `(model + tools + infrastructure + review + retries) / accepted tasks`.
- Throughput-only duration estimate: `requested output tokens / measured output tokens per second`.

Terminology

Open weights means downloadable parameters under stated terms; it does not necessarily mean an OSI-approved license, open training data or unrestricted commercial use. **Context window** is the maximum supported token capacity, not guaranteed effective recall. **Reasoning effort** is an inference control that can alter quality, latency and cost. **Independent** means the measurement was produced by an operator other than the model vendor; it does not mean error-free.

Change triggers

Refresh this paper when the model identifier, provider route, index version, price, license, system card or release status changes. Do not silently update one number: a new evidence cutoff should produce a new edition and rerun the peer table.

References

1. **[S01]** Artificial Analysis (2026). Artificial Analysis Intelligence Benchmarking Methodology. Evidence tier T2.
2. **[S02]** Artificial Analysis (2026). Artificial Analysis model leaderboard. Evidence tier T2.
3. **[S03]** Terminal-Bench (2026). terminal-bench@2.1 leaderboard. Evidence tier T2.
4. **[S04]** SWE-bench Team (2026). SWE-bench official leaderboards. Evidence tier T2.
5. **[S05]** Elham Tabassi, NIST (2023). Artificial Intelligence Risk Management Framework 1.0. Evidence tier T1.
6. **[S06]** Chloe Autio et al., NIST (2024). Artificial Intelligence Risk Management Framework: Generative AI Profile. Evidence tier T1.
7. **[S07]** Long Phan et al. (2026). Humanity's Last Exam. Evidence tier T1.
8. **[S08]** xAI (2026). Introducing Grok 4.6. Evidence tier T3.
9. **[S09]** xAI (2026). Grok 4.6 model documentation. Evidence tier T3.
10. **[S10]** xAI (2026). xAI API. Evidence tier T3.
11. **[S11]** xAI (2026). xAI safety. Evidence tier T3.
12. **[S12]** Artificial Analysis (2026). Grok 4.6 benchmarks and analysis. Evidence tier T2.