AI INFRASTRUCTURE · OCTOBER 8, 2026

Google Gemini 4 Argon: reported strengths and weaknesses, as of 8 October 2026

Gemini 4 Argon, announced 30 Sep 2026, leads many knowledge-work, long-context and multimodal benchmarks and posts a low hallucination rate. It trails rivals on terminal agent tasks, some coding and computer-use tests, and verbosity. Most scores are Google's own, and access is limited to Fairwind partners.

9 min read · 12 sources

2026 | AI safety | API pricing | Gemini 4 Argon | Google DeepMind | benchmarks | cybersecurity

CONTENTS

What Argon is and who can use it 2
Reported strengths 3
- Knowledge work (legal, finance, business automation) 3
- Long context and multimodal 3
- Coding 3
- Cybersecurity 4
- Hallucination, prompt injection and preference rankings 4

Reported weaknesses 4
- Not the leader on terminal and some coding tasks 4
- Lower accuracy, verbosity and cost 5
- Mixed legal result, and speed is unknown 6
- Availability, ecosystem and product trust 6

How far to trust the numbers 7
Safety posture 7
Bottom line 8
Sources 8

"Argon" is Gemini 4 Argon, Google DeepMind's frontier model, announced on 30 September 2026. Google positions it for real-world software engineering, enterprise knowledge work such as legal and finance, and defensive cybersecurity[1][2]. This note collects what Google, independent evaluators and community commentators have reported about where it is strong and where it is not. Almost nobody outside a small partner group has used it yet, so the evidence is mostly benchmarks, not hands-on reviews.

77.9%	68.9%	53	$2 / $10	15%
DeepSWE v1.1 (Google, self-computed)	Vals Index, #1 of the models listed	Artificial Analysis Intelligence Index, #8 of 225	Intro price per 1M in/out tokens ($4 / $20 after)	AA-Omniscience hallucination rate (vs 51% GPT-6 Astra)

What Argon is and who can use it

Argon is a proprietary reasoning model that takes text and image input, outputs text, and has a **1M-token context window** [4]. Google also raised the output limit from 64K to **1M tokens**. Two details qualify that claim. Vals lists a 262K maximum output, and Artificial Analysis reached 1M only through a new "Long Decode Continuation" API feature that pauses and resumes long responses across calls [5].

Access is the biggest practical limit. Argon is rolling out only to trusted cyber defenders in Google's Fairwind Program, plus government users [1] [5]. Google says it will refine guardrails before releasing to developers, enterprises and consumers "as soon as possible", starting with paid API customers and Google AI Ultra subscribers. It gives no date [1]. Google says it is also taking part in the US government's voluntary pre-release access process [1]. A third-party developer guide, last updated 2 October, found no confirmed public Gemini API model ID or pricing-page entry [3]. I did not re-check Google's public pricing page myself.

Fairwind itself is tightly controlled. Google says it works with over 650 partners and runs background checks on applicants. Partners may only grant access to internal cybersecurity, incident-response or penetration-testing teams, and may not share or resell access [9]. Permitted dual-use work includes authorised threat simulation, reverse engineering and malware analysis, while creating malware is barred [9]. Argon supports zero data retention when accessed as a managed model on Gemini Enterprise [9]. Academic labs doing defensive benchmarking can apply, but ordinary developers and students are pointed to CodeMender with publicly available models [9].

Pricing is **$2 per million input tokens and $10 per million output tokens** as an "introductory" rate. The footnote says that after the introductory period the price becomes **$4 and $20** [1]. No end date has been announced [5]. Cached input is discounted 95% [1].

- 2 Sep 2026 — **Fairwind Program launched**
  Limited-access cyber defense program, initially built on Gemini 3.8 Flash Cyber
- 30 Sep 2026 — **Gemini 4 Argon announced**
  Fairwind partners and government users only
- 1 Oct 2026 — **Independent evaluations appear**
  Artificial Analysis, Vals and Arena publish results

Not announced

Broad release

Paid API customers and Google AI Ultra first, per Google

Reported strengths

Knowledge work (legal, finance, business automation)

This is where Google's claims are strongest. In Google's own comparison, Argon leads all four of the Knowledge-work rows against GPT-6 Astra, Claude Fable 5.1 and Claude Opus 5.5[2]:

BENCHMARK	ARGON	GPT-6 ASTRA	FABLE 5.1	OPUS 5.5
Vals Index	68.9%	63.1%	65.8%	67.0%
AutomationBench	51.3%	41.4%	31.4%	42.5%
Vals Finance Agent v2	65.4%	53.5%	58.9%	58.6%
Harvey's Legal Agent Benchmark	19.6%	5.4%	6.7%	3.8%

Vals's own leaderboard confirms Argon at #1 on the Vals Index. It scores 68.90% at **$15.68 per test**, ahead of Claude Sonnet 5.5 (67.04%, $21.34) and Opus 5.5 (66.97%, $32.14), and finishes in 46m33s against about 1h18m for the Claude models[6]. The Vals table lists Argon at $4/$20 per million tokens, the standard rate, so that cost-per-test figure is not a discounted one. Vals's takeaways add that no single model leads every component, and that the top three Claude models' scores include some tasks served by a fallback model after a provider refusal[6]. A Latent Space roundup also reports that Argon uses roughly a quarter of Sonnet 5.5's output tokens on Vals Index tasks[5]. Artificial Analysis separately ranks Argon #1 on its AutomationBench-AA at 77.5%[5].

Long context and multimodal

Google's table shows the widest margins on long context and video:

- **GraphWalks, 256K to 1M tokens (BFS F1):** 84.2% against 71.8% for Astra, 66.8% for Opus 5.5 and 65.0% for Fable 5.1. Up to 128K, Argon scores 99.7%[2].
- **LVBench (long-video understanding):** 91.7% against 87.5% for Astra[2].
- **Chartography:** 71.6% against 71.0% for Astra and 46.2% for Fable 5.1[2].
- **LABBench2 and RiemannBench (science and math):** 88.8% and 76.0%, both ahead of every comparator[2].
- **Agent's Last Exam (computer use):** 39.5% against 38.2% for Opus 5.5 and 34.2% for Astra[2].

Coding

Argon sets a reported state of the art on **DeepSWE v1.1 at 77.9%**, against 74.2% for Opus 5.5 and 74.1% for Astra. It also leads Vibe Code Bench at 91.9%[2]. Vals reports that Argon built 30 Vibe Code

Bench apps perfectly, against 25 for Opus 5 and 24 for Astra[5]. PostTrainBench shows 45.3%, up from 21.99% for Gemini 3.1 Pro, though Opus 5.5 is higher at 49.3%[2][5].

Google also describes internal use. Argon agents reportedly freed over 300 TiB of data-center memory (an estimated 500 TiB to 1 PiB in total). They are migrating C/C++ to Rust at scale, up to 800K+ lines for the Fuchsia Zircon kernel. In libgav1 they replaced 32K lines of SIMD code and produced a decoder that runs 2.7x faster than the earlier Rust port with identical output[1]. Google also says Argon beat a published quantum-subroutine baseline by 40% in minutes[1]. These are Google's own accounts and are not independently verified. Google notes the rewrites are still undergoing auditing and review before production[1].

Cybersecurity

Google says Argon can "autonomously find, validate, and patch critical software vulnerabilities"[1]. On CWE-bench v1 it ties for first at **68%**. The Fairwind page chart shows Grok 4.7, Argon and GPT-6 Astra all at 68%, with Opus 5.5 at 67%[9]. So the lead is a tie, not a win. Google says Wiz used it to find a critical flaw in healthcare software used by hospitals worldwide that earlier frontier models had missed[1].

The larger gains are against Google's own earlier cyber model. Per chart descriptions on the Fairwind page, Argon scores 85.8% on an internal real-world vulnerability discovery set against 71.0% for Gemini 3.8 Flash Cyber. On Wiz's penetration-test benchmark it scores 70.9% against 58.2%[9]. Both are internal benchmarks without external replication. Vals reports 70% on CyberBench proof-of-concept tasks[5]. Defenders in the program get a version without cyber guardrails[1].

Hallucination, prompt injection and preference rankings

Per a social-media summary relayed by Latent Space, Argon's hallucination rate on AA-Omniscience is **15%, against 51% for Astra**[5]. On Gray Swan's indirect prompt-injection benchmark, Argon has the lowest attack success rate shown, 0.7% at k=15 attempts, against 52.7% for Kimi K3 at the other end of the chart[9]. Argon is #1 in Text Arena at 1525 and #1 for steerability in a preliminary Agent Arena sample of about 3K sessions[5]. Artificial Analysis rates Argon's cost per Intelligence Index task at $1.99 against $3.26 for Astra, at the discounted price[4][5].

Reported weaknesses

Not the leader on terminal and some coding tasks

Argon's results on terminal-style agent work and some coding suites are middling or behind:

Terminal-Bench 4.0 (Google's table)



Source: DeepMind evaluation table [2]. Argon's score is self-computed.

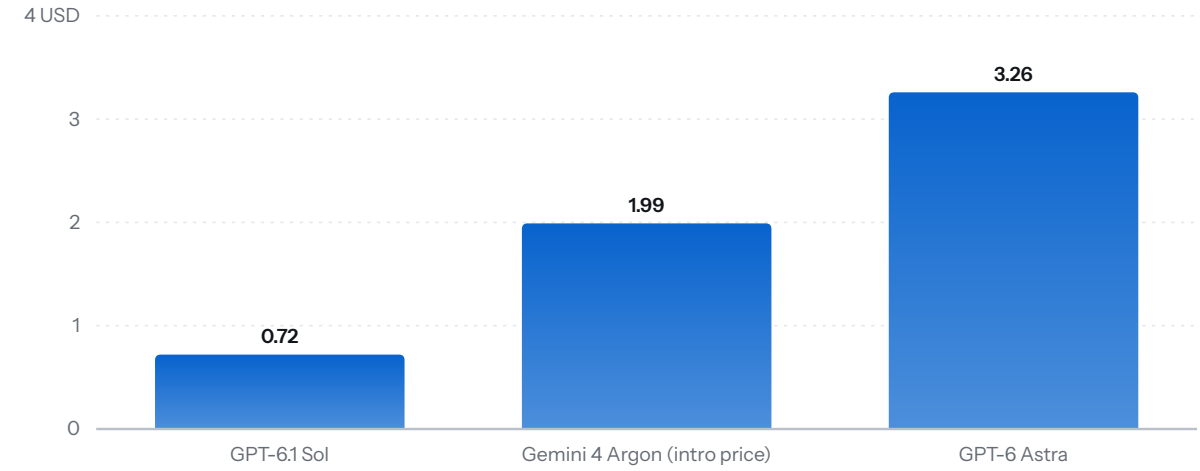
Beyond Terminal-Bench 4.0, the same table shows Argon behind on [2]:

- **FrontierSWE v2:** 55.0% against 65.5% for Astra and 62.3% for Opus 5.5.
- **Terminal-Bench Science 0.1:** 57.6% against 68.1% for Astra.
- **OSWorld-2.0 (computer use):** 69.2% against 72.6% for Astra.
- **PostTrainBench:** 45.3% against 49.3% for Opus 5.5.

Arena places Argon only #8 on Code Arena WebDev (1679) and #8 overall in Agent Arena [5]. Google's launch post shows only the favourable charts and never mentions these losses [1].

Lower accuracy, verbosity and cost

The low hallucination rate comes with a trade-off. On AA-Omniscience, accuracy is reported at **50% against Astra's 63%**, which suggests Argon declines or abstains more often [5]. On overall index score, Argon's 53 only ties Astra and ranks #8 of 225 [4] [5]. Argon used **110M output tokens** on the index, which Artificial Analysis calls "somewhat verbose" against a median of 81M [4]. Latent Space reports about 62K output tokens per task against 27K for Astra and notes the cost advantage "comes from price, not efficiency". At standard pricing the per-task cost would be about $3.98 [5].



Source: Artificial Analysis page and Latent Space [4] [5].

In a Hacker News thread on the analysis, a commenter pointed out that GPT-6.1 Sol scores one index point lower (52) at $0.72 per task [7]. Argon is not the cheapest route to near-frontier scores. At the $4/$20 standard price its per-token cost would equal Claude Opus 5.5's in Vals's table [6].

Mixed legal result, and speed is unknown

Argon's 19.6% on Harvey's Legal Agent Benchmark leads Google's comparison but trails Muse Spark's listed 25.42% in one critic's post [5]. Vals notes that its best component score is only 23.75%, so there is a lot of headroom [6]. Artificial Analysis shows Argon's output speed as "N/A" [4], so latency is unmeasured.

Availability, ecosystem and product trust

Most people cannot test Argon, and the developer guide notes that the launch thread is split on availability, benchmark selectivity and Google product continuity [3]. The huge Hacker News launch thread (1,702 points, about 1,190 comments a week after launch) is largely about Google's Antigravity ("agy") harness and Ultra subscriptions, not about Argon's output, because few commenters have the model [10]. Complaints there include being forced into Google's harness, and fear of account bans without appeal. These are user opinions, not findings. Others praise Gemini 3.8 Flash and the harness, which says more about Google's ecosystem than about Argon [10]. One commenter recalled Gemini 3.0 as benchmarking well but needing "constant babysitting" for agentic work [7]. That is anecdotal, but it shows the scepticism about launch numbers.

How far to trust the numbers

Google's methodology page says Argon scores are pass@1, run with the highest thinking settings. Non-Gemini scores come mostly from providers' self-reported numbers [8]. Several key Argon results were self-computed by Google. These include DeepSWE v1.1, Terminal-Bench 4.0, PostTrainBench, LABBench2, GraphWalks, Agent's Last Exam, OSWorld-2.0 and LVBench [8]. Some comparisons are uneven. For LVBench, Google used 1 FPS for Gemini, 800 frames for GPT-6 Astra, 300 for Fable 5.1 and 600 for Opus 5.5, citing API limits [8]. For OSWorld-2.0, Argon's number is the best of three runs and uses a compaction procedure borrowed from the Opus 5.5 system card [8]. On Agent's Last Exam and OSWorld-2.0, Fable 5.1 or Anthropic results are missing [2] [8]. The cyber results that Google calls "internal" run in an internal Antigravity harness that is not cyber-specialised [8].

> CAUTION
>
> The sources disagree on Terminal-Bench. Google's table lists Terminal-Bench 4.0 at 57.4% and Terminal-Bench Science at 57.6% [2]. The Latent Space digest cites 57% for Terminal-Bench 4 and a Vals post attributing 57.6% to Terminal-Bench 4.0, up from 19.0% [5]. I could not resolve this from the primary sources.

Independent confirmation exists for some results (Vals Index, Artificial Analysis index, Arena) but those leaderboards measure different things [4] [5] [6]. Critics on social media alleged preference-data "benchmaxxing" and objected to the DeepSWE figures [5]. These are opinions, not demonstrated findings. TechCrunch's launch coverage also notes that Google's claim of beating Astra, Fable and Opus rests on its own blog's benchmark selection [11]. The NYT article on the launch could not be retrieved, so I did not use it.

Safety posture

Google lists four areas of safeguards: misuse defense (including cyber and CBRN refusals plus monitoring of internal activations), prompt-injection resistance, monitoring of chain-of-thought and actions to stop out-of-bounds behaviour, and hardened sandboxes [1]. It calls Argon its "most resilient model yet" against indirect prompt injection and says it leads Gray Swan's IPI benchmark [1]. It says training runs were monitored by a similar system, with findings deliberately not fed back into training so as not to teach the model to evade monitoring [1]. These are Google's statements, and I found no independent audit. The offer of an unguarded cyber variant to vetted defenders is itself a risk trade-off Google has chosen [1]. Fairwind's governance rules (authentication, phishing-resistant MFA, no resale) are the main control on that variant [9].

Bottom line

On published evidence Argon is a real frontier entrant. It is strongest at professional knowledge work, long-context retrieval, video and chart understanding, vulnerability work and calibrated (low-hallucination) answers. It is weakest at terminal-style agent runs, some SWE suites, computer use and token efficiency, and the discounted price hides a verbosity cost. Most headline numbers are Google's own, and almost no one outside Fairwind can test it. Treat it as a watchlist model until public API access, final pricing and independent real-workload results arrive, and run your own evaluations when it does.

Terms used in this document (4)

- **benchmaxxing** — Tuning a model to score well on public benchmarks at the expense of real-world performance.
- **Fairwind Program** — Google DeepMind's limited-access programme giving vetted governments, infrastructure operators and technology platforms early access to its cyber-defence models.
- **Hallucination rate** — The share of answers in which a model asserts something false instead of abstaining or answering correctly.
- **pass@1** — A benchmark score from a single attempt per problem, with no majority voting or parallel test-time compute.

SOURCES

1 Gemini 4 Argon: our next era of frontier intelligence (Google blog)
https://blog.google/innovation-and-ai/models-and-research/gemini-models/gemini-4-argon/

2 Gemini — Google DeepMind (Argon performance table)
https://deepmind.google/models/gemini/

3 Gemini 4 Argon: What Developers Can Actually Use (Developers Digest)
https://www.developersdigest.tech/blog/gemini-4-argon-developer-guide-2026

4 Gemini 4 Argon (High) analysis — Artificial Analysis
https://artificialanalysis.ai/models/gemini-4-argon

5 [AINews] Gemini 4 Argon: GDM's answer to Astra/Fable, with 1M output (Latent Space)
https://www.latent.space/p/ainews-gemini-4-argon-gdms-answer

6 Vals Index Leaderboard and Methodology
https://www.vals.ai/benchmarks/vals_index

7 Hacker News: Gemini 4 Argon (High) analysis thread
https://news.ycombinator.com/item?id=49914236

8 Gemini 4 Argon model evaluation methodology (DeepMind)
https://deepmind.google/models/evals-methodology/gemini-4-argon

9 Fairwind Program — Google DeepMind
https://deepmind.google/fairwind-program/

10 Hacker News: Gemini 4 Argon launch thread
https://news.ycombinator.com/item?id=49913571

11 Google releases Gemini 4 Argon, called its most powerful model yet (TechCrunch)
https://techcrunch.com/2026/09/30/google-releases-gemini-4-argon-called-its-most-powerful-model-yet/

12 Proactive cyber defense for governments and enterprises (Google blog, Fairwind launch)
https://blog.google/innovation-and-ai/technology/safety-security/fairwind-program/

Researched and written with claude-sonnet-5-5, from the sources listed above. Figures are as reported by those sources. Exported November 29, 2026 from https://intel.innerpulse.net/doc/google-gemini-4-argon-reported-strengths-and-weaknesses-as-of-8-october-2026, where the current version lives.