Gemini 3.8 Flash: Accuracy, Latency, and Cost for Research Tasks

IntuitionLabs.ai · Adrien Laurent · 2026-09-05 · gemini 3.8 flash · google gemini · llm benchmarks · ai api pricing · llm for research · artificial analysis · gemini flash pricing · ai model review 2026

Executive Summary

Google DeepMind released **Gemini 3.8 Flash** on September 2, 2026, marking the third Flash-tier launch in roughly six weeks after Gemini 3.6 Flash (July 21, 2026) and Gemini 3.7 Flash (August 13, 2026) (9to5google.com) (9to5google.com). The model launched generally available (GA), not in preview, with a release date of September 2, 2026 (docs.cloud.google.com), and with a 1,048,576-token (roughly 1 million) input context window, a 64,000-token output ceiling, and support for text, image, audio, and video inputs (deepmind.google) (deepmind.google). Google priced application programming interface (API) access at an introductory $0.75 per 1 million input tokens and $3.75 per 1 million output tokens, rising to $1.50 and $7.50 respectively once the promotion expires on December 31, 2026 (blog.google) (blog.google).

On research-relevant benchmarks, the independent evaluation platform Artificial Analysis measured Gemini 3.8 Flash at 59 points on its Intelligence Index at high reasoning effort, a 3-point gain over Gemini 3.7 Flash (artificialanalysis.ai), alongside an output speed near 300 to 305 tokens per second (artificialanalysis.ai).

Independent Vals AI test runs, aggregated by the benchmark tracker BenchLM, put Gemini 3.8 Flash at 80.0% on SWE-bench and 81.3% on Terminal-Bench 2.1, both meaningfully below Google's own self-reported 89.4% to 90.8% on the same terminal benchmark (benchlm.ai) (benchlm.ai) (docs.cloud.google.com).

For research-specific tasks, Gemini 3.8 Flash posted the best published score on LABBench2, a real-world biology research benchmark, at 86.2%, ahead of Claude Opus 5's 84.2% (vellum.ai), and its largest gain on the harder half of BioMysteryBench, reaching 56.5% against Claude Opus 5's 49.4% (vellum.ai). On document and table extraction it trailed rather than led: 35.0% on the GDP.PDF comprehension benchmark against 40.0% for GPT-5.6 Sol (vellum.ai). Cost efficiency is real but qualified: Artificial Analysis found the model uses roughly 30% more output tokens per task than Gemini 3.7 Flash, which raised its per-task cost to $0.58 at high effort even as it remained the cheapest model at its intelligence tier (artificialanalysis.ai) (artificialanalysis.ai). This report, prepared by the life sciences and artificial intelligence (AI) consultancy **IntuitionLabs**, synthesizes these vendor and independent sources into a reproducible, task-level view of accuracy, latency, and cost for research use, rather than a single universal ranking (intuitionlabs.ai).

Introduction and Background

Large language models (LLMs) are increasingly used as research assistants: summarizing literature, extracting structured data from tables and PDFs (portable document format files), and writing analysis code. Google DeepMind's **Gemini 3.8 Flash**, published on September 2, 2026, is the newest entrant in this category, positioned by Google as the primary "workhorse" of its Gemini 3 model family for agentic and reasoning-heavy work (deepmind.google) (docs.cloud.google.com). Gemini 3.8 Flash is an iteration on Gemini 3.7 Flash rather than a new base architecture (deepmind.google), released alongside a separately distributed, cybersecurity-focused sibling, Gemini 3.8 Flash Cyber, that is out of scope for this review because it targets vulnerability discovery rather than general research work.

Scope and method. This report does not run a proprietary benchmark suite against Gemini 3.8 Flash; no such claim is made anywhere below. Instead, it synthesizes three categories of evidence collected and verified during the week of September 5, 2026: Google's own model card, developer documentation, and launch announcement (vendor claims); independent evaluation platforms including Artificial Analysis, BenchLM (which aggregates third-party Vals AI test runs), and Vellum's own analysis of the launch benchmark table; and developer community reports from Hacker News, labeled explicitly as sentiment rather than measurement. Every quantitative claim below identifies the category of source from which it came.

This distinction matters most for teams evaluating **large language models (LLMs)** for research tasks such as literature synthesis, structured data extraction from scientific PDFs, and scientific or statistical code generation, where source-grounding accuracy and reproducibility carry more weight than raw chat-style benchmark performance. IntuitionLabs previously published a broader cross-vendor pricing comparison covering four API providers, updated through February 2026 (intuitionlabs.ai); this report does not repeat that comparison and instead focuses narrowly on what changed with a single new model release and how that change should be measured for research use.

Product Overview: Architecture, Availability, and Pricing

Gemini 3.8 Flash accepts text, image, audio, and video input and returns text-only output, with a **context window** (the maximum combined input plus conversation history a model can process at once) of up to 1,048,576 tokens and a maximum output of 64,000 tokens (deepmind.google) (ai.google.dev). Its knowledge cutoff, the date after which the model has no training data, is March 2026 (deepmind.google). The model exposes three configurable **thinking levels**, LOW, MEDIUM, and HIGH, defaulting to MEDIUM, which trade added reasoning steps for higher accuracy and cost (docs.cloud.google.com). Google's own developer documentation frames the design change plainly: at higher effort, the model "takes smaller reasoning steps, calls tools iteratively, and verifies its work along the way," rather than answering in a single pass (ai.google.dev). It is now the default model behind Google's Antigravity agent and Antigravity software development kit (SDK), Google's agent-building surfaces (ai.google.dev).

Table 1 summarizes the model's core specifications as documented by Google as of September 2026.

Attribute	Gemini 3.8 Flash
Release date / launch stage	September 2, 2026; generally available (GA), not preview (docs.cloud.google.com)
Input context window	Up to 1,048,576 tokens (roughly 1 million) (deepmind.google)
Max output	64,000 tokens (deepmind.google)
Input / output modalities	Text, image, audio, video in; text only out (deepmind.google)
Thinking levels	LOW / MEDIUM (default) / HIGH (docs.cloud.google.com)
Knowledge cutoff	March 2026 (deepmind.google)
Base lineage	Iteration on Gemini 3.7 Flash, same family (deepmind.google)
Restricted sibling model	Gemini 3.8 Flash Cyber, distributed only through Google's Fairwind Program to vetted defenders (blog.google)

The table shows a model built for long documents and multi-step agent work rather than for the largest possible single output. Its 1-million-token context window can hold very long source documents (an entire clinical study report or a large codebase, for instance), while the 64,000-token output ceiling constrains how much a single response, such as a fully rewritten long document, can return in one call.

Google's own developer guide describes Gemini 3.8 Flash Cyber, the vulnerability-focused sibling, as reaching "a success rate exceeding 70%" on an internal real-world vulnerability-discovery benchmark, and 47.2% pass@1 on the external CWE-Bench (Common Weakness Enumeration benchmark) patching test, versus 47.8% for an unnamed leading frontier model at lower cost (blog.google). Because that variant is not broadly available and does not target research workflows, this report does not evaluate it further. On safety, Google's Frontier Safety Framework assessment found Gemini 3.8 Flash reaches no new tracked or critical capability levels relative to Gemini 3.7 Flash (deepmind.google), though the model card notes non-English safety performance "regressed slightly relative to 3.7 Flash" (deepmind.google), and Google's own card acknowledges the model "may exhibit some of the general limitations of foundation models, such as hallucinations" (deepmind.google), a limitation directly relevant to source-grounded research summarization, addressed further below.

Performance and Benchmark Analysis

Google's own developer guide reports Gemini 3.8 Flash scoring 90.8% on Terminal-Bench 2.1 (an agentic terminal-use benchmark), up from 81.6% for Gemini 3.7 Flash, and 61.6% on SWE-Bench Pro versus 60.4% previously (docs.cloud.google.com) (docs.cloud.google.com).

Independent test runs diverge further. BenchLM's tracker, which aggregates results run by the third-party evaluator Vals AI, puts Gemini 3.8 Flash at 81.3% on Terminal-Bench 2.1 and explicitly labels Google's own 89.4% figure a "provider run," distinct from an independently verified score (benchlm.ai) (benchlm.ai). The same Vals AI runs give Gemini 3.8 Flash 80.0% on SWE-bench, 94.4% on GPQA Diamond (Graduate-Level Google-Proof Question and Answering, a hard science multiple-choice benchmark, versus 95.5% for Gemini 3.1 Pro as the best verified score), and 90.2% on MMLU-Pro (Massive Multitask Language Understanding, versus 92.4% for Claude Fable 5.1) (benchlm.ai) (benchlm.ai) (benchlm.ai). Across these three independently run benchmarks, Gemini 3.8 Flash trails the best verified competitor score by roughly 1 to 17 percentage points, a pattern worth weighing against any single vendor-reported win.

Not every comparison favors incumbents. On Humanity's Last Exam Verified (HLE-Verified, a multidisciplinary expert-reasoning test), Gemini 3.8 Flash posts the best figure in Google's own table at 54.9%, ahead of GPT-5.6 Sol's 54.5% and Claude Opus 5's 54.4%, a genuinely close three-way result (vellum.ai); the score is otherwise sourced only to Google's own model card, per BenchLM (benchlm.ai). Conversely, on Terminal-bench 4.0, a broader agent-work benchmark, Gemini 3.8 Flash scores 19.1% against Claude Opus 5's 51.8%, per the same Google-published table (vellum.ai).

On throughput, Artificial Analysis measured roughly 300 to 305 output tokens per second at high reasoning effort, the fastest speed the platform says it had measured to that point, alongside an average 2.5-minute completion time per task (artificialanalysis.ai) (vellum.ai). Separately, BenchLM recorded the same 327 tokens-per-second output rate against a field median of only 92 tokens per second (benchlm.ai), On raw latency, BenchLM separately recorded a 10.75-second time to first token (TTFT), the delay before any output begins, alongside a 327 tokens-per-second output rate once generation starts (benchlm.ai); a model can be fast once it starts responding and still feel slow to a user waiting for the first word, particularly at higher thinking levels.

Table 2 consolidates the benchmark figures gathered above, distinguishing which organization produced each score.

Benchmark	Gemini 3.8 Flash	Comparator	Source type
Terminal-Bench 2.1	90.8% (dev guide) / 89.4% (blog)	Gemini 3.7 Flash: 81.6%; Claude Opus 5: 89.1%	Vendor self-reported (docs.cloud.google.com) (vellum.ai)
Terminal-Bench 2.1 (Vals AI run)	81.3%	Best verified: GPT-6 Astra, 87.3%	Independent (benchlm.ai)
Terminal-bench 4.0	19.1%	Claude Opus 5: 51.8%	Vendor self-reported (vellum.ai)
SWE-bench (Vals AI run)	80.0%	Claude Opus 5 (best verified): 97.0%	Independent (benchlm.ai)

Benchmark	Gemini 3.8 Flash	Comparator	Source type
HLE-Verified	54.9%	GPT-5.6 Sol: 54.5%; Claude Opus 5: 54.4%	Vendor self-reported (Google model card) (vellum.ai)
GPQA Diamond (Vals AI run)	94.4%	Gemini 3.1 Pro (best verified): 95.5%	Independent (benchlm.ai)
MMLU-Pro (Vals AI run)	90.2%	Claude Fable 5.1 (best verified): 92.4%	Independent (benchlm.ai)
CharXiv Reasoning (charts, no tools)	86.2%	GPT-5.6 Terra: 85.9%; Claude Opus 5: 83.7%	Vendor self-reported (vellum.ai)
GDP.PDF (document comprehension)	35.0%	GPT-5.6 Sol: 40.0%; Claude Opus 5: 37.0%	Vendor self-reported (vellum.ai)
BioMysteryBench (hard tier)	56.5%	Claude Opus 5: 49.4%; GPT-5.6 Sol: 44.7%	Vendor self-reported (vellum.ai)
LABBench2 (biology research tasks)	86.2% (best published)	Claude Opus 5: 84.2%	Vendor self-reported (vellum.ai)

Table 2 shows a model that leads on reasoning-heavy, self-reported scientific and agentic benchmarks (HLE-Verified, BioMysteryBench, LABBench2, CharXiv), sits several points behind the field on independently run general-coding benchmarks (SWE-bench, Terminal-Bench 2.1 via Vals AI), and trails badly on the broadest long-horizon agent test (Terminal-bench 4.0). No single row supports a universal "best model" claim; the pattern instead argues for matching the specific task type to the specific evidence above rather than to a single headline score.

Independent benchmark firm Artificial Analysis's public-facing article separately reported an overall Intelligence Index rank of 17th out of 196 tracked models, against a class median of 36, using its own aggregated methodology distinct from the per-model detail-page figures discussed above (eesel.ai); BenchLM's separate, differently weighted composite score for the same model is 78.41 out of 100, ranking it 6th of 232 models it tracks (benchlm.ai). These two independent trackers do not use the same weighting and should not be treated as interchangeable confirmations of one another.

Accuracy for Research Tasks: Grounding, Extraction, and Scientific Code

Research work depends less on chat fluency than on three specific capabilities: source-grounded summarization that does not hallucinate facts not present in the source, accurate extraction of tables and figures from PDFs, and correct generation of scientific or statistical code. Evidence on the first is thin. Google's own model card states plainly that the model "may exhibit some of the general limitations of foundation models, such as hallucinations" (deepmind.google), and no named independent organization has yet published a dedicated faithfulness or groundedness score (a measure of whether a summary states only what the source document actually says) specific to this model; that absence is recorded here as a genuine gap rather than filled with an estimate.

On document and table extraction, the evidence is more concrete, and it argues against treating Gemini 3.8 Flash as uniformly best-in-class. Google's document-processing documentation states the model can "analyze and interpret content, including text, images, diagrams, charts, and tables, even in long documents up to 1000 pages" and can "extract information into structured output formats" such as JavaScript Object Notation (JSON) (ai.google.dev) (ai.google.dev). Yet on the specific GDP.PDF benchmark discussed above, an "all-pass" measure of expert document comprehension, Gemini 3.8 Flash scored 35.0%, behind both GPT-5.6 Sol (40.0%) and Claude Opus 5 (37.0%), per Google's own published table (vellum.ai). On chart and figure reasoning specifically (CharXiv), by contrast, it led the field at 86.2% (vellum.ai). Practically, this suggests the model is stronger at reasoning over a chart it can already see clearly than at exhaustively pulling every field out of a dense, multi-page PDF, a distinction worth testing directly on a given team's own document types before deployment.

Scientific coding and bioinformatics evidence is stronger and more favorable. Google's evaluation methodology document describes how its BioMysteryBench and LABBench2 benchmarks are run: "models are given access to a Linux terminal with pre-installed bioinfo tools, Python, and R, as well as internet access," restricted to an allow-listed domain set defined by each benchmark's authors (deepmind.google), meaning these are agentic, tool-using evaluations rather than static question-answering tests. LABBench2 itself is a composite of eleven sub-tasks, including image-based and PDF-based table question-answering sub-benchmarks (TableQA2-img, TableQA2-pdf) directly relevant to literature-derived data extraction, plus a clinical-trial question-answering sub-task, TrialQA (deepmind.google). On the composite LABBench2 score, Gemini 3.8 Flash posted the best published figure at 86.2%, ahead of Claude Opus 5's 84.2% (vellum.ai), and on the harder half of BioMysteryBench it improved most sharply of any model tested, moving from Gemini 3.7 Flash's 43.5% to 56.5%, ahead of Claude Opus 5 (49.4%) and GPT-5.6 Sol (44.7%) (vellum.ai). Google's own long-context evaluation, GDM-MRCR v2, reports results as a cumulative score at a 128,000-token depth, described in the methodology document as "self computed" rather than independently audited (deepmind.google).

Tool integration extends beyond the model itself. Google DeepMind's own "Science Skills" repository, installable as agent tooling inside the Antigravity environment where Gemini 3.8 Flash is now the default model, bundles "agent skills for scientific research tasks, spanning genomics, structural biology, cheminformatics, literature search, and more" across more than 30 integrated databases and tools (raw.githubusercontent.com). This is a meaningful capability signal for research teams building agent workflows, though it describes tooling availability rather than an independently verified accuracy score, and no public source located during this research documents a specific pilot use of Gemini 3.8 Flash inside a pharmaceutical or regulatory document workflow; that too is recorded as a gap rather than an inferred fact.

Reception and Community Perspectives

The developer sentiment below is reported as community perception, not as benchmarked measurement, and should be weighted accordingly. On Hacker News, one commenter (simonw) described the model favorably: "the speed combined with the fact that this thing is really good at HTML JavaScript is pretty exciting," reporting a demo built for roughly 1.8 cents in 13 seconds (news.ycombinator.com). Another commenter (colechristensen) called the prior Flash release "competitive with opus/fable and also FAST," suggesting Google's Flash line is underrated relative to larger competing models (news.ycombinator.com).

Other developers pushed back specifically on the token-efficiency point raised in the Data Analysis section below. One Hacker News commenter (bermudi) noted the model "putting it dead last in output tokens per task in the leaderboard," despite its raw speed advantage (news.ycombinator.com). A separate thread compared Gemini's usefulness as a plan-reviewing agent against Claude and a GPT-based coding agent (Codex) and reported that "Claude could always note much more problems in Codex's plan and implementation than Gemini could," estimating the gap at roughly ten to one in Claude's favor for that specific reviewing task (news.ycombinator.com). Relayed via a third-party review site, another Hacker News commenter (markasoftware) argued that once effort levels are matched, "it's only equal to opus 5 medium effort. Opus 5 max scores 63," a more measured framing than the vendor's own headline comparisons (eesel.ai). A Google developer advocate, quoted in the same third-party analysis, characterized the token-usage increase as an intentional design tradeoff: "it takes smaller steps and verifies its work more often" (eesel.ai). At the more critical end, one social media post described the model as "absolutely not usable for SWE" (software engineering) for the poster's own workload, and a further commenter (WhitneyLand) flagged that "it's not even close to Opus 5 on Terminal-bench 4.0, 19.1% vs. 51.8%, (eesel.ai) (eesel.ai).

Taken together, community reaction mirrors the benchmark split documented above: enthusiasm for raw speed and web-style coding, tempered skepticism about token efficiency and about the model's standing on the hardest, longest-horizon agent tasks. None of the sources reviewed for this section describe any safety incident, data-integrity failure, or business dispute tied to the model; the discussion is limited to task performance and cost.

Data Analysis and Evidence

Google's introductory pricing, effective at launch and holding through December 31, 2026, is $0.75 per 1 million input tokens and $3.75 per 1 million output tokens, identical to Gemini 3.7 Flash's introductory rate (apidog.com), independently transcribed from Google's pricing page by the tracking site TokenCost on September 3, 2026 (tokencost.app). Context caching, a feature that lets repeated large-document prompts reuse prior processing at a discount, carries an introductory rate of $0.075 per 1 million tokens through the same date (ai.google.dev), and batch processing, for non-time-sensitive workloads, carries a 50% cost reduction versus standard pricing (ai.google.dev). Separately, Google Cloud is crediting 50% of net Provisioned Throughput spending on Gemini 3.8 Flash, 3.7 Flash, and 3.6 Flash from August 13 through December 31, 2026 (cloud.google.com). All of these promotional rates expire on the same calendar date regardless of a given model's launch date, and every rate doubles on January 1, 2027 to $1.50 input and $7.50 output per 1 million tokens (datacamp.com); one third-party cost tracker warned bluntly that "if you are sizing a 2027 budget on $0.75, you are sizing it on the wrong number" (eesel.ai).

Independent analysis firm Apidog confirmed that Gemini 3.8 Flash's introductory per-token pricing is identical, row for row, to Gemini 3.7 Flash's, tracing both back to the same rows on Google's own pricing page (apidog.com) (apidog.com), with the same $0.075 cached-context rate and 50% batch discount carried over unchanged (apidog.com) (apidog.com). Identical per-token pricing does not mean identical real-world cost, however: Apidog, citing Artificial Analysis, reports that Gemini 3.8 Flash used roughly 30% more output tokens per completed task than Gemini 3.7 Flash, averaging 48,000 tokens per task (apidog.com), a finding corroborated separately by MindStudio's own pricing analysis, which notes that "Flash-family models can use up to 30% more output tokens per task than the previous generation, partially offsetting the savings" of a lower sticker price (mindstudio.ai).

Artificial Analysis's own headline figures show the model at $0.58 per Intelligence Index task at high reasoning effort, still the cheapest model at its intelligence tier (artificialanalysis.ai), and one review reported that running Artificial Analysis's entire evaluation suite against the model cost $825.83 in total API spend (eesel.ai).

Table 3 places Gemini 3.8 Flash's per-token pricing alongside publicly reported competitor rates as of September 2026; because two secondary sources gave different figures for the same competitor model, both are shown rather than silently reconciled.

Model	Input $ / 1M tokens	Output $ / 1M tokens	Note
Gemini 3.8 Flash (introductory, through Dec 31, 2026)	$0.75	$3.75	Official Google pricing, confirmed independently (apidog.com)
Gemini 3.8 Flash (standard, from Jan 1, 2027)	$1.50	$7.50	Official Google pricing, confirmed independently (datacamp.com)
GPT-5.6 Sol	$5 (DataCamp) / $4, promo through late Nov (Coursiv)	$30 (DataCamp) / $20 (Coursiv)	Two secondary sources disagree; neither traces to an OpenAI primary source (datacamp.com) (coursiv.io)
Claude Opus 5	Reported as $5.25 blended (MindStudio); not separately confirmed input/output	Reported as $5.25 blended	Different pricing methodology from other rows; shown for directional context only (mindstudio.ai)
Claude Fable 5.1	$10	$50	Per DataCamp (datacamp.com)

Even accounting for the discrepancies in Table 3, every competitor figure reported by any source sits well above Gemini 3.8 Flash's introductory rate; Coursiv's own summary put the gap at "roughly a fifth of Sol's price and a sixth or seventh of Opus 5's" per token (coursiv.io). The three points that matter for a research budget are, first, that this per-token gap is real and large regardless of which competitor figure is used; second, that the 30% higher output-token usage documented above erodes but does not eliminate that gap; and third, that every number in this section is promotional and will roughly double on January 1, 2027, so any multi-year cost projection needs to be built on the standard rate, not the launch-week rate.

Implications and Future Directions

Three Flash-tier releases in six weeks (Gemini 3.6, 3.7, and 3.8 Flash between July 21 and September 2, 2026) signal that Google is iterating its mid-tier model on a roughly three-week cadence rather than the multi-quarter cycle common for frontier-tier models (9to5google.com) (9to5google.com). For a research team, that cadence has a practical consequence: benchmark comparisons and cost models built around any single Flash-tier snapshot have a shelf life measured in weeks, not years, and should be re-run rather than assumed durable. General "best model" roundups for research work tend to change from month to month and rarely isolate narrow, tool-using science tasks specifically; the bioinformatics-specific benchmark data gathered in this report (LABBench2, BioMysteryBench) argues for evaluating research-task fit at that narrower level, where Gemini 3.8 Flash now posts the best published figures against Claude Opus 5 specifically.

For teams choosing between Google's Flash and Pro tiers, the single clearest available data point is GPQA Diamond, where an independent Vals AI run gave Gemini 3.1 Pro 95.5% against Gemini 3.8 Flash's 94.4%, a narrow one-point gap despite Flash's dramatically lower per-token price and higher throughput (benchlm.ai). No source reviewed for this report published a broader, task-matched Flash-versus-Pro comparison; teams needing a definitive answer for their own workload should treat that as a genuine open question rather than assume Flash is a strict subset of Pro's accuracy.

As an adjacent life sciences and artificial intelligence consultancy rather than a model vendor, IntuitionLabs' role in this landscape is advisory: helping research and regulatory-adjacent teams translate benchmark tables like those above into task-specific pilot plans, particularly where document extraction accuracy or reasoning-cost tradeoffs carry compliance weight (intuitionlabs.ai). Readers looking for a broader, multi-vendor pricing comparison across chat and API tiers, rather than this report's narrower focus on one model's research-task evidence, can consult IntuitionLabs' earlier comparison of four leading provider pricing structures (intuitionlabs.ai). The most durable practical guidance from the evidence above is procedural rather than a model pick: pilot on a representative sample of a team's own documents and code before committing budget.

Frequently Asked Questions (FAQs)

Is Gemini 3.8 Flash more accurate than Gemini Pro for research tasks?

The only independently run, task-matched comparison available shows Gemini 3.1 Pro slightly ahead on GPQA Diamond, a graduate-level science benchmark, 95.5% versus 94.4% (benchlm.ai). No comprehensive independent Flash-versus-Pro comparison across research-specific tasks was found; this report treats that as an open question.

What does Gemini 3.8 Flash cost per token?

$0.75 per 1 million input tokens and $3.75 per 1 million output tokens introductory, through December 31, 2026, then $1.50 and $7.50 respectively (apidog.com) (datacamp.com), though real per-task cost runs higher because the model uses roughly 30% more output tokens per task than its predecessor (apidog.com).

How fast is Gemini 3.8 Flash?

Independent measurements cluster around 300 to 327 output tokens per second once generation begins (benchlm.ai), but one tracker separately recorded a 10.75-second delay before the first token appears, meaning the perceived responsiveness at the start of a request is slower than the sustained throughput number suggests (benchlm.ai).

Is Gemini 3.8 Flash good for life sciences and biology research?

On the two benchmarks specifically designed for real-world biology and bioinformatics research tasks (LABBench2 and BioMysteryBench), it posted the best published figures against Claude Opus 5 and GPT-5.6 Sol (vellum.ai) (vellum.ai), and Google DeepMind's own Science Skills tooling bundles genomics, structural biology, and literature-search integrations for use with the model (raw.githubusercontent.com). No public source documents a specific pharmaceutical or regulatory pilot deployment.

Research-use considerations

Document and table extraction from dense PDFs, where it scored behind two competitors on the GDP.PDF benchmark despite leading on chart-based reasoning (vellum.ai), and the absence of any published, independent source-grounding or hallucination-rate benchmark specific to this model.

Which organizations report benchmark results?

Artificial Analysis and Vals AI runs aggregated by BenchLM report benchmark results cited throughout this report.

Conclusion

Gemini 3.8 Flash is a genuine, generally available upgrade to Google's mid-tier Flash line, not a preview or an announced-only product, launched September 2, 2026 at unchanged introductory pricing versus its predecessor. The evidence gathered here does not support a single verdict on whether it is the "best" model for research tasks; it supports a task-specific one. It leads on biology-research and finance and legal reasoning benchmarks and on raw output speed, while trailing on independently run coding benchmarks, on the broadest long-horizon agent test, and on dense document extraction. Its real per-task cost runs above its sticker price because of higher token usage per task, and every currently quoted price is promotional through the end of 2026. Any team adopting Gemini 3.8 Flash for research work should treat the tables above as a starting hypothesis to test against its own documents and code, not as a substitute for that test.
