Google Gemini 3.8 Flash Brings Better Coding and Agent Performance at the Same Token Price

TechRounder PDF Edition

Live article: https://www.techrounder.com/ai/google-gemini-3-8-flash-brings-better-coding-and-agent-performance/

By Vipin PG | Published September 3, 2026 | Updated September 3, 2026 | Format: Analysis | 6 min read

In brief

Google released Gemini 3.8 Flash on September 2, 2026, only three weeks after Gemini 3.7 Flash. It is the third Flash release from Google in roughly six weeks.

Google released Gemini 3.8 Flash on September 2, 2026, only three weeks after Gemini 3.7 Flash. It is the third Flash release from Google in roughly six weeks.

The new model focuses heavily on coding, AI agents and jobs that require the model to work through several steps before producing a result. Google has also introduced Gemini 3.8 Flash Cyber, a separate security-focused version with restricted access.

Gemini 3.8 Flash keeps the introductory API pricing of 3.7 Flash. The actual cost of running it can still be higher because the new model tends to spend more tokens on difficult tasks.

For developers and businesses using Gemini regularly, that difference matters more than the version number.

What Google released

The main release is Gemini 3.8 Flash, with the model ID:

‘gemini-3.8-flash‘

Google positions it for software development, AI agents and multi-step professional work.

The company also announced Gemini 3.8 Flash Cyber. It uses the same underlying model family but has additional training for defensive cybersecurity work such as finding vulnerabilities and helping create patches.

Gemini 3.8 Flash Cyber is available through Google’s Fairwind Program rather than the normal public API. Access is limited to approved organisations such as government agencies, critical-infrastructure operators and software maintainers.

Most users will therefore be dealing with the regular Gemini 3.8 Flash model.

Gemini 3.8 Flash specifications

The model has a context window of 1,048,576 input tokens and can generate up to 65,536 output tokens.

It accepts:

- Text
- Images
- Video
- Audio

- PDF files

Output is currently text only.

Gemini 3.8 Flash also works with several tools available through Google's Gemini platform, including function calling, code execution, Google Search grounding, Maps grounding, URL context, file search, structured output and computer use in preview.

Google provides three thinking levels:

- 'low'
- 'medium'
- 'high'

Medium is the default.

The 'minimal' thinking setting supported by some earlier Gemini models is not available with Gemini 3.8 Flash. Requests using that setting return an error.

Gemini 3.8 Flash API pricing

Google is keeping Gemini 3.8 Flash at the same introductory API rate as Gemini 3.7 Flash until December 31, 2026.

API usage: Input | Price per 1 million tokens: $0.75

API usage: Output, including thinking tokens | Price per 1 million tokens: $3.75

API usage: Cached input | Price per 1 million tokens: $0.075

Google's published pricing changes from January 1, 2027:

API usage: Input | Price per 1 million tokens: $1.50

API usage: Output, including thinking tokens | Price per 1 million tokens: $7.50

API usage: Cached input | Price per 1 million tokens: $0.15

For developers in India, the useful number is not only the price per million tokens. Token consumption across thousands or millions of requests has a direct effect on the monthly cost.

Gemini 3.8 Flash appears willing to spend more tokens when a task needs deeper reasoning.

Why the same API rate can still cost more

Artificial Analysis measured Gemini 3.8 Flash at 59 on its Intelligence Index using high reasoning. Gemini 3.7 Flash scored 56 under the same comparison.

The newer model also cost more per benchmark task.

Artificial Analysis estimated about $0.58 per task for Gemini 3.8 Flash at high reasoning, compared with roughly $0.40 for Gemini 3.7 Flash.

The main reason was greater token usage. Its testing found that Gemini 3.8 Flash produced roughly 30% more output tokens and took additional turns during agent-style tasks.

Reasoning level made a large difference:

- Low reasoning: about $0.24 per task
- Medium reasoning: about $0.41
- High reasoning: about $0.58

These figures come from benchmark testing rather than Google's API rate card. Real costs will depend on the workload.

They still reveal an important behaviour change. Gemini 3.8 Flash can spend more time working through a difficult problem, and that extra work consumes tokens.

For routine use, medium reasoning is likely to make more sense than leaving every request on high.

Coding appears to be the biggest improvement

The clearest gains are showing up in coding and agent tasks.

These are jobs where the model may need to read several files, inspect existing code, use tools, correct mistakes and continue working before it can finish.

On DeepSWE v1.1, a benchmark for longer software-engineering tasks, Gemini 3.8 Flash scored 73.7% in Google's comparison.

Gemini 3.7 Flash scored 65.3%.

Google's table placed Claude Opus 5 at 74.0% and GPT-5.6 Sol at 72.7% on the same evaluation.

That puts Gemini 3.8 Flash unusually close to larger models on this particular type of coding work.

Terminal-Bench 2.1 showed a similar result. Gemini 3.8 Flash reached 89.4%, compared with 85.8% for Gemini 3.7 Flash.

Google also reported gains on finance, legal and computer-use evaluations.

The pattern matters more than any single benchmark score. Gemini 3.8 Flash appears to perform best when a task has a clear objective and the model can keep working through multiple steps to reach it.

Bigger frontier models still have advantages

Other benchmarks show a wider gap.

On Terminal-Bench 4.0, Gemini 3.8 Flash scored 19.1%. Gemini 3.7 Flash scored 11.2%, so there was a clear improvement.

Claude Opus 5 reached 51.8% on Google's comparison, while GPT-5.6 Sol reached 37.3%.

Gemini 3.8 Flash also remained behind larger models on parts of Google's broader computer-use and professional-work evaluations.

This makes workload selection important.

A defined coding task inside a repository gives the model clear boundaries and a measurable goal. A long autonomous task can require much more planning, judgement and recovery when something goes wrong.

Gemini 3.8 Flash has become much stronger at the first type of work. Developers should still test harder autonomous workflows carefully before replacing the larger model already handling them.

What early users are reporting

Public testing started quickly after the release, especially among people using Gemini through Antigravity and coding workflows.

Several early reports describe 3.8 Flash as more patient than 3.7 Flash.

Gemini 3.7 Flash had a reputation among some users for reaching an answer too quickly during coding tasks. Testers using 3.8 Flash on high reasoning reported more checking and a greater willingness to stay with a problem.

There are still limits.

One detailed Antigravity review described Gemini 3.8 Flash as much better than 3.7 while finding GPT-5.6 Sol more persistent on difficult long-running work.

Other users reported good results on short and medium coding jobs, including web development and single-file tasks.

Higher token consumption also came up repeatedly in early feedback. A model that spends longer checking its work can use more of a user’s available quota.

These reports are early observations from individual users. They are useful for spotting patterns, but they should not carry the same weight as controlled testing.

Gemini 3.8 Flash Cyber has a different purpose

Gemini 3.8 Flash Cyber is one of the more unusual parts of the release.

Google built it for defensive cybersecurity work, with extra focus on vulnerability discovery and automated patching.

On CWE-Bench testing listed in the original evaluation material, Gemini 3.8 Flash Cyber reached 47.2% pass@1 using a high-thinking Antigravity setup.

Google also reported an 86.2% pass@1 result on CyberGym and 71% on an internal vulnerability-discovery test covering 20 programming languages.

Some of the Cyber results come directly from Google or organisations working with Google. They should be read as vendor or partner results rather than independent measurements.

Access is another important restriction.

Gemini 3.8 Flash Cyber is being provided through the Fairwind Program to vetted users. Ordinary Gemini API developers should not expect to see it as another model in the public model list.

Where Gemini 3.8 Flash is available

Google announced Gemini 3.8 Flash across several products.

Developers can use it through services including:

- Gemini API
- Google AI Studio
- Google Antigravity

Google has also announced the model for Gemini Enterprise and other supported developer products.

For consumers, access includes Google AI Pro and Ultra users in supported Gemini experiences.

The release may not appear in every Gemini account at the same moment. Early reports showed differences between products and accounts during the initial rollout.

The model available through the Gemini API and the model currently selected inside the Gemini consumer app should therefore be checked separately.

Should you move from Gemini 3.7 Flash to 3.8 Flash?

For coding and agent work, Gemini 3.8 Flash is the more interesting option.

Users working with multi-file coding, longer development tasks, document-heavy analysis or tool-based agents have the most to gain from the new model.

Gemini 3.7 Flash can still make sense for simple workloads where speed, predictable behaviour and lower token use matter more than additional reasoning.

Businesses processing large request volumes should pay particular attention to this difference. A small increase in tokens per request becomes significant at scale.

Medium reasoning is a sensible starting point for Gemini 3.8 Flash. High reasoning can be reserved for work where the extra effort produces a useful improvement.

Teams currently using larger frontier models should test the same real workload on both models before switching. Benchmark rankings can point to strengths, but production tasks often expose differences that a benchmark cannot show.

The Flash model is becoming more capable

Google’s recent release pattern shows how much attention it is giving the Flash family.

Gemini 3.8 Flash arrived only three weeks after 3.7 Flash. The main gains are concentrated in areas that developers regularly spend money on: coding, tool use and multi-step agent work.

The cost side needs equal attention. Gemini 3.8 Flash can use more tokens to produce those better results, especially at high reasoning.

For a developer choosing between 3.7 and 3.8 today, I would start with Gemini 3.8 Flash on medium reasoning and run it against a real set of production tasks. Keep 3.7 Flash available for simpler, high-volume jobs where its lower token use still saves money.