Gemini 3 Flash: frontier intelligence built for speed

Gemini 3 Flash is our latest model with frontier intelligence built for speed that helps everyone learn, build, and plan anything —faster.

Figure 1: Gemini 3 Flash text

Today, we're expanding the Gemini 3 model family with the release of Gemini 3 Flash, which offers frontier intelligence built for speed at a fraction of the cost. With this release, we're making Gemini 3's next-generation intelligence accessible to everyone across Google products.

Last month, we kicked off Gemini 3 with Gemini 3 Pro and Gemini 3 Deep Think mode, and the response has been incredible. Since launch day, we have been processing over 1T tokens per day on our API. We've seen you use Gemini 3 to vibe code simulations to learn about complex topics, build and design interactive games and understand all types of multimodal content.

With Gemini 3, we introduced frontier performance across complex reasoning, multimodal and vision understanding and agentic and vibe coding tasks. Gemini 3 Flash retains this foundation, combining Gemini 3' s Pro-grade reasoning with Flash-level latency, efficiency and cost. It not only enables everyday tasks with improved reasoning, but also is our most impressive model for agentic workflows.

Starting today, Gemini 3 Flash is rolling out to millions of people globally:

- For developers in the Gemini API in Google AI Studio, Gemini CLI and our new agentic development platform Google Antigravity
- For everyone via the Gemini app and in AI Mode in Search
- For enterprises in Vertex AI and Gemini Enterprise

Gemini 3 Flash: frontier intelligence at scale

Gemini 3 Flash demonstrates that speed and scale don't have to come at the cost of intelligence. It delivers frontier performance on PhD-level reasoning and knowledge benchmarks like GPQA Diamond (90.4%) and Humanity's Last Exam (33.7% without tools), rivaling larger frontier models, and significantly outperforming even the best 2.5 model, Gemini 2.5 Pro, across a number of benchmarks. It also reaches state-of-the-art performance with an impressive score of 81.2% on MMMU Pro, comparable to Gemini 3 Pro.

In addition to its frontier-level reasoning and multimodal capabilities, Gemini 3 Flash was built to be highly efficient, pushing the Pareto frontier of quality vs. cost and speed. When processing at the highest thinking level, Gemini 3 Flash is able to modulate how much it thinks. It may think longer for more complex use cases, but it also uses 30% fewer tokens on average than 2.5 Pro, as measured on typical traffic, to accurately complete everyday tasks with higher performance.

Gemini 3 Flash pushes the Pareto frontier on performance vs. cost and speed.

Performance, here, is measured by LMArena Elo Score.

Gemini 3 Flash's strength lies in its raw speed, building on the Flash series that developers and consumers already love. It outperforms 2.5 Pro while being 3x faster (based on Artificial Analysis benchmarking) at a fraction of the cost. Gemini 3 Flash is priced at $0.50/1M input tokens and $3/1M output tokens (audio input remains at $1/1M input tokens).

Gemini 3 Flash outperforms 2.5 Pro in speed and quality.

For developers: intelligence that keeps up

Gemini 3 Flash is made for iterative development, offering Gemini 3's Pro-grade coding performance with low latency —it's able to reason and solve tasks quickly in high-frequency workflows. On SWE-bench Verified, a benchmark for evaluating coding agent capabilities, Gemini 3 Flash achieves a score of 78%, outperforming not only the 2.5 series, but also Gemini 3 Pro. It strikes an ideal balance for agentic coding, production-ready systems and responsive interactive applications.

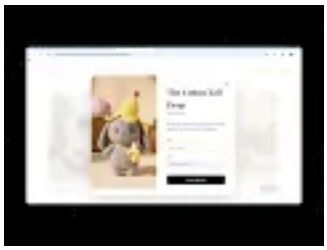


Figure 2: Demo of Gemini 3 Flash for developers

Gemini 3 Flash in Google Antigravity works quickly to update production-ready applications.

Gemini 3 Flash's strong performance in reasoning, tool use and multimodal capabilities is ideal for developers looking to do more complex video analysis, data extraction and visual Q&A, which means it can enable more intelligent applications —like in-game assistants or A/B test experiments —that demand both quick answers and deep reasoning.

Gemini 3 Flash enables multimodal reasoning in a hand-tracked "ball launching puzzle game"game providing near real-time AI assistance.

Gemini 3 Flash builds and A/B tests new loading spinner designs in near real-time, streamlining the design-to-code process.

Gemini 3 Flash uses multimodal reasoning to analyze and caption an image with contextual UI overlays in near real-time, ultimately transforming a static image into an interactive experience.

Gemini 3 Flash takes a single instruction prompt and codes three unique design variations.

We've received a tremendous response from companies using Gemini 3 Flash. Companies like JetBrains, Bridgewater Associates, and Figma are already using it to transform their businesses, recognizing how its inference speed, efficiency and reasoning capabilities perform on par with larger models. Gemini 3 Flash is available today to enterprises via Vertex AI and Gemini Enterprise.

"In our JetBrains AI Chat and Junie agentic-coding evaluation, Gemini 3 Flash delivered quality close to Gemini 3 Pro, while offering significantly lower inference latency and cost. In a quota-constrained production setup, it consistently stays within per-customer credit budgets, allowing complex multi-step agents to remain fast, predictable, and scalable."

Denis Shiryaev
Head of AI DevTools Ecosystem @ JetBrains

"At Bridgewater, we require models capable of reasoning over vast, unstructured multimodal datasets without sacrificing conceptual understanding. Gemini 3 Flash is the first to deliver Pro-class depth at the speed and scale our workflows demand. Its long-context performance on complex problems is exceptional."

Jasjeet Sekhon
Chief Scientist and Head of AI @ AIA Labs, Bridgewater Associates

“Gemini 3 Flash is a great option for teams who want to quickly test and iterate on product ideas in Figma Make. The model can rapidly and reliably create prototypes while maintaining attention to detail and responding to specific design direction.”

Loredana Crisan
Chief Design Officer @ Figma

“Our engineers have found Gemini 3 Flash to work well together with Debug Mode in Cursor. Flash is fast and accurate at investigating issues and finding the root cause of bugs.”

Lee Robinson
VP of Developer Experience @ Cursor

“Gemini 3 Flash remains the best fit for Warp’s Suggested Code Diffs, where low latency and cost efficiency are hard constraints. With this release, it resolves a broader set of common command-line errors while staying fast and economical. In our internal evaluations, we’ve seen an 8% lift in fix accuracy.”

Zach Lloyd
Founder & CEO @ Warp

“Gemini 3 Flash has achieved a meaningful step up in reasoning, improving over 7% on Harvey's BigLaw Bench from its predecessor, Gemini 2.5 Flash. These quality improvements, combined with Flash's low latency, are impactful for high-volume legal tasks such as extracting defined terms and cross-references from contracts.”

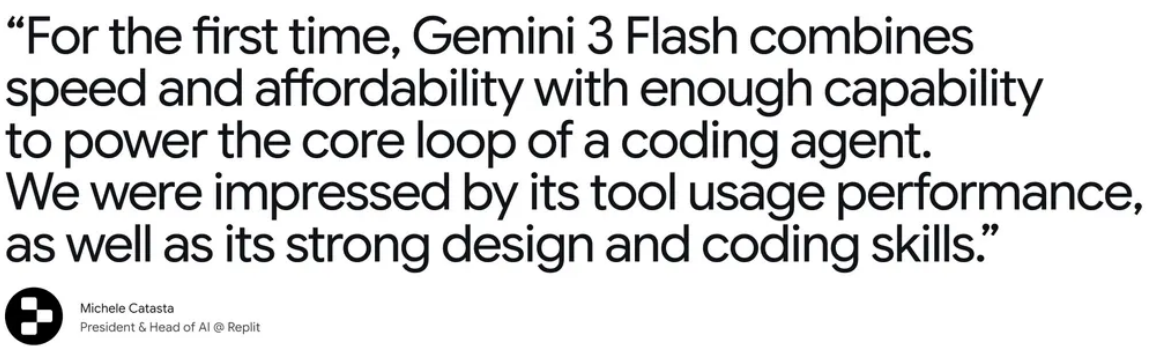
Niko Grupen
Head of Applied Research @ Harvey

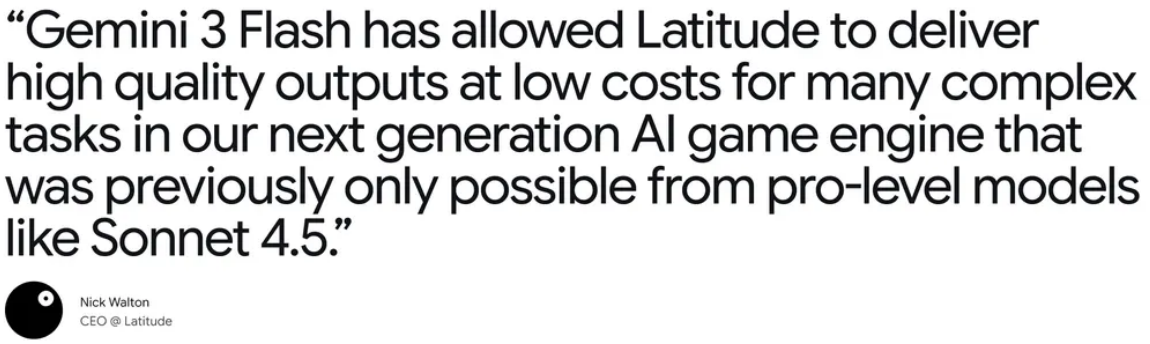
“Astrocade is using Gemini 3 Flash for our agentic game creation engine to power coding and planning. The speed of the 3 Flash model allows us to generate full game-level plans from a single prompt, but with decreased latency allowing us to deliver fast responses for our users.”

Ali Sadeghian
Co-Founder & CTO @ Astrocade

“Presentations.ai is using Gemini 3 Flash to enhance our intelligent slide-generation agents, and we’re consistently impressed by the pro-level quality at lightning-fast speeds. With previous pro sized models there were many things we simply couldn’t attempt because of the speed vs. quality tradeoff. With Gemini 3 Flash, we’re finally able to explore those workflows.”

Saravanan Govindaraj
Co-Founder & Head of Product Development @ Presentations.AI





For everyone: Gemini 3 Flash is rolling out globally

Gemini 3 Flash is now the default model in the Gemini app, replacing 2.5 Flash. That means all of our Gemini users globally will get access to the Gemini 3 experience at no cost, giving their everyday tasks a major upgrade.

Because of Gemini 3 Flash’s incredible multimodal reasoning capabilities, you can use it to help you see, hear and understand any type of information faster. For example, you can ask Gemini to understand your videos and images and turn that content into a helpful and actionable plan in just a few seconds.

Gemini 3 Flash in the Gemini app can analyze short video content and give you a plan, like how to improve your golf swing.

As Gemini 3 Flash is optimized for speed, it can see and guess what you're drawing while you're still sketching it.

You can upload an audio recording and Gemini 3 Flash will identify your knowledge gaps, create a custom quiz, and give you detailed explanations on the answers.

Or you can quickly build fun, useful apps from scratch using your voice without prior coding knowledge. Just dictate to Gemini on the go, and it can transform your unstructured thoughts into a functioning app in minutes.

Figure 3: Food prototype using Gemini 3 Flash

Describe an idea using Gemini 3 Flash and turn it into a working prototype in minutes.

Gemini 3 Flash is also starting to roll out as the default model for AI Mode in Search with access to everyone around the world.

Building on the reasoning capabilities of Gemini 3 Pro, AI Mode with Gemini 3 Flash is more powerful at parsing the nuances of your question. It considers each aspect of your query to serve thoughtful, comprehensive responses that are visually digestible —pulling real-time local information and helpful links from across the web. The result effectively combines research with immediate action: you get an intelligently organized breakdown alongside specific recommendations —at the speed of Search.

This shines when tackling complex goals with multiple considerations like trying to plan a last-minute trip or learning complex educational concepts quickly.

Figure 4: Demo of Gemini 3 Flash in AI Mode

Gemini 3 Flash brings the incredible reasoning capabilities of Gemini 3 to Search, without compromising speed, so you can tackle your most complicated questions.

Try Gemini 3 Flash today

Gemini 3 Flash is available now in preview via the Gemini API in Google AI Studio, Google Antigravity, Vertex AI and Gemini Enterprise. You can also access it through other developer tools like Gemini CLI and Android Studio. It's also starting to roll out to everyone in the Gemini app and AI Mode in Search, bringing fast access to next-generation intelligence at no cost.

We're looking forward to seeing what you bring to life with this expanded family of models: Gemini 3 Pro, Gemini 3 Deep Think and now, Gemini 3 Flash.