

[Back to Models](#)

Google: Gemini 3.8 Flash

Active

50% OFF

google/gemini-3.8-flash

Compare

API Request

Chat

Gemini 3.8 Flash is Google's most intelligent workhorse model yet, delivering major gains over 3.7 Flash in software engineering, agentic workflows, and multi-step reasoning for specialized domains. It approaches the quality of higher-cost frontier models while still offering configurable effort levels for balancing performance and token efficiency.

By: Google Input Type: Output Type: Publish time: 2026-09-02

Providers

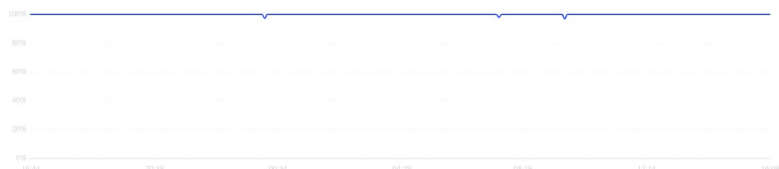
Route requests across multiple providers. Copy a provider slug to set your preference.

Provider	Cache Hit Rate	Input	Output	Cache read	Context	Latency	Throughput
Google Vertex google-vertex	87.7%	\$1.5 \$0.75 / M tokens	\$2.5 \$3.75 / M tokens	Read: \$1.5 0.075 / M tokens Write: - / M tokens	1.05M	5.21s	109tps

Uptime

24hours

Direct request success rate on AI Gateway and per-provider.



Sep 18, 16:09

Google Vertex

100.00%

ZenMux Launches Gemini 3.8 Flash Model Access for Agentic Engineering Workflows

October 01, 2026

Singapore - October 01, 2026 -

ZenMux has added Gemini 3.8 Flash to its model-routing platform through a Google Vertex provider route. The dedicated listing records a publish date of September 2, 2026 and a context window of 1.05 million tokens. Google positions Gemini 3.8 Flash as a workhorse-tier model, not a flagship reasoning model, designed for high-volume, everyday engineering and agentic tasks. ZenMux describes the model as an option for software engineering, agentic workflows, and multistep reasoning where teams need to balance model capability with token efficiency.

The launch adds another Google model to the ZenMux catalog and gives developers a consistent place to review its identifier, context specification, provider route, and current platform measurements. The listing characterizes Gemini 3.8 Flash as a workhorse model with configurable effort levels. That control is intended to let application teams tune how much reasoning effort is applied to a request, subject to the parameters and behavior supported by the active provider route.

The ZenMux Gemini 3.8 Flash page identifies software engineering and agent-based work among the model's principal use cases. In a coding environment, teams may evaluate the model for tasks such as explaining a repository, drafting changes, reviewing errors, or coordinating several tool-assisted steps. Results can vary across languages, codebases, prompts, and test conditions, so the listing should be treated as a starting point for evaluation rather than a guarantee of performance.

ZenMux lists Google Vertex as the current provider route for the model. Provider routing through a unified catalog can simplify how an application team adds a model to an existing evaluation framework. The surrounding application can retain common logging, error handling, access controls, and model-selection logic while developers assess the behavior of the newly listed option. Google Vertex availability, regional support, quotas, and other provider conditions may still apply.

The 1.05-million-token context window creates room for workloads involving long documents, extensive conversation history, multiple code files, or collections of source material. Effective use of that capacity depends on how information is selected and structured. Large requests can include duplicated or irrelevant content, and they may require additional processing time. Teams can use retrieval, preprocessing, prompt organization, and output limits to keep requests aligned with the task.

The ZenMux page currently displays route pricing, cache information, response measurements, and throughput observations. These values can change with provider updates, traffic, caching behavior, and the time at which measurements are taken. The launch announcement does not present any dynamic metric as a fixed service level.

Teams reviewing ZenMux Gemini 3.8 Flash can use the dedicated page to confirm the current model identifier and Google Vertex route. They can then compare the model with other catalog options using representative prompts, quality checks, latency measurements, and cost controls. Human review remains appropriate for consequential outputs and for tasks where factual or code errors could create material risk.

The Gemini 3.8 Flash addition supports ZenMux's continuing work to expand model choice through a unified routing platform. The company organizes technical specifications and provider information so development teams can move from discovery to structured evaluation without maintaining a separate catalog for each provider.

For more information, visit <https://zenmux.ai/google/gemini-3.8-flash>

About ZenMux:

ZenMux is an enterprise-grade large model aggregation platform with an insurance payout mechanism. The platform provides one-stop access to the latest models across providers. When issues such as poor output quality or excessive latency occur during use, our intelligent insurance detection and payout mechanism automatically compensates, addressing enterprise concerns around AI hallucinations and unstable quality.

###

For more information about ZenMux, contact the company here: ZenMuxEmberember@zenmux.ai Singapore

ZenMux

ZenMux is an enterprise-grade AI model aggregation platform with a built-in insurance payout mechanism, offering one-stop access across providers and automatic compensation for quality or latency issues.

Website: <https://zenmux.ai/>

Email: ember@zenmux.ai

AI

Simplified & Assured.

ZenMux