Gemini 3.1 Flash-Lite: Google's Fastest Model Costs $0.25 per Million Tokens

Kabui, Charles

2026-03-06

Read at ToKnow.ai

Google released Gemini 3.1 Flash-Lite, the cheapest and fastest model in its Gemini 3 series. It costs **$0.25 per million input tokens** and **$1.50 per million output tokens**, roughly half the price of Gemini 3 Flash ($0.50/$3.00). On Artificial Analysis benchmarks, it outputs at **288 tokens per second** with a blended cost of $0.56 per million tokens, making it faster than Grok 4.1 Fast (121 tok/s at $0.28) and significantly cheaper than Claude 4.5 Haiku ($2.00).

The model accepts text, images, video, audio, and PDFs across a **1 million token context window** with up to 64K output tokens. It supports thinking (with configurable levels from minimal to high), function calling, structured JSON output, code execution, search grounding, and context caching. Google describes it as their "workhorse model built for cost-efficiency and high-volume tasks." The Gemini CLI already uses Flash-Lite as a classifier to route queries to Flash or Pro based on task complexity.

The target audience is clear: developers running high-volume pipelines where per-token cost is the bottleneck. Translation at scale, document triage, entity extraction, transcription, and lightweight agent routing are the use cases Google highlights. A free tier is available for prototyping, with paid access for production. Context caching drops the input cost further to $0.025 per million tokens for repeated content. For comparison, Qwen3.5-35B-A3B, Alibaba's unified multimodal agent model, offers a similar efficiency play (3B active of 35B total parameters) with open weights and 201-language support, though it targets a different deployment model: self-hosted rather than API-first.

The competitive floor for "good enough" AI keeps dropping. A model with a 1 million token context window, multimodal input, and built-in reasoning now costs less than a quarter per million tokens. The open question is whether intelligence scores, where Flash-Lite trails GPT-5 mini and Claude 4.5 Haiku on leaderboards, will matter more than speed and price for the bulk of production workloads.

Sources:

- Gemini 3.1 Flash-Lite Model Card
- Gemini 3 Developer Guide
- Gemini API Pricing
- Artificial Analysis LLM Leaderboard
