

Gemini 3.8 Flash

Google

AT A GLANCE

PRICE PER 1M

\$0.825 / \$4.125

input / output · cached \$0.083

CONTEXT

1.05M

tokens in

MAX OUTPUT

66K

tokens per reply

TAKES

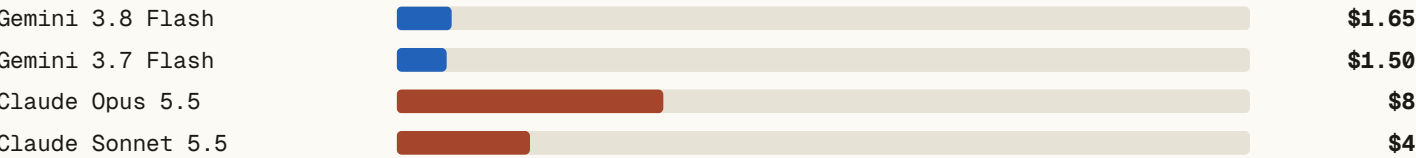
text, image, video, file, audio

returns text

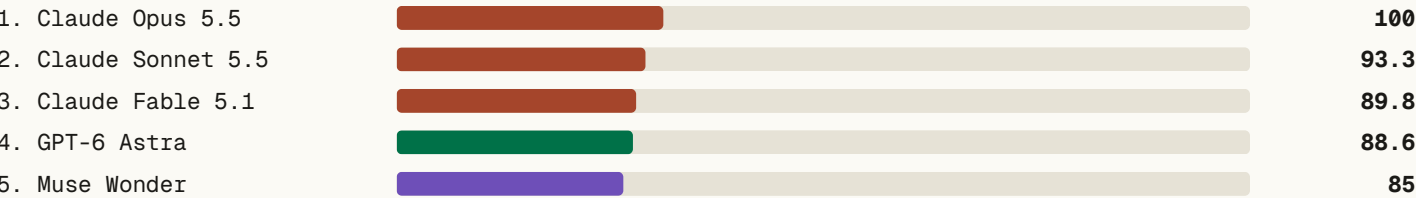
HOW IT COMPARES

MODEL	ROLE	IN / 1M	OUT / 1M	CONTEXT	BENCH INDEX
Gemini 3.8 Flash	new	\$0.825	\$4.125	1.05M	46.4
Gemini 3.7 Flash	it follows	\$0.75	\$3.75	1.05M	—
Claude Opus 5.5	today's #1	\$4	\$20	1M	100
Claude Sonnet 5.5	best value	\$2	\$10	1M	93.3
vs Gemini 3.7 Flash		10% more	the model it follows: \$0.75 / \$3.75		
vs Claude Opus 5.5		79% cheaper	today's #1: \$4 / \$20, bench index 100; context 1M		
vs Claude Sonnet 5.5		59% cheaper	best value: \$2 / \$10, bench index 93.3; context 1M		

PRICE PER 1M TOKENS, BLENDED 3 IN : 1 OUT



WHERE THE PUBLIC LEADERBOARDS STAND TODAY



Gemini 3.8 Flash sits at 46.4 on the bench index. The bench index is the median of each model's percentile across the public leaderboards that rate it (LMArena, LiveBench, SimpleBench, LLM Stats, Vals AI, Artificial Analysis).

Sources: price, limits and modalities for Gemini 3.8 Flash as the OpenRouter catalog lists google/gemini-3.8-flash; the other models as their makers list them (models.dev); rankings from the ArtificialWatch Bench Index. Facts only: no forecast of ours, no claims from the lab's announcement.