

GLM-5.3 Flash

Zhipu AI

AT A GLANCE

PRICE PER 1M

\$0.15 / \$0.50

input / output · cached
\$0.03

CONTEXT

1.05M

tokens in

MAX OUTPUT

944K

tokens per reply

TAKES

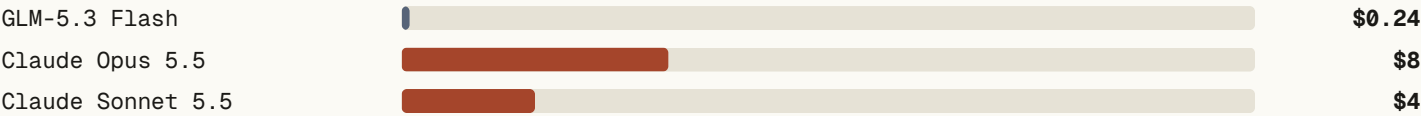
text, image, video

returns text

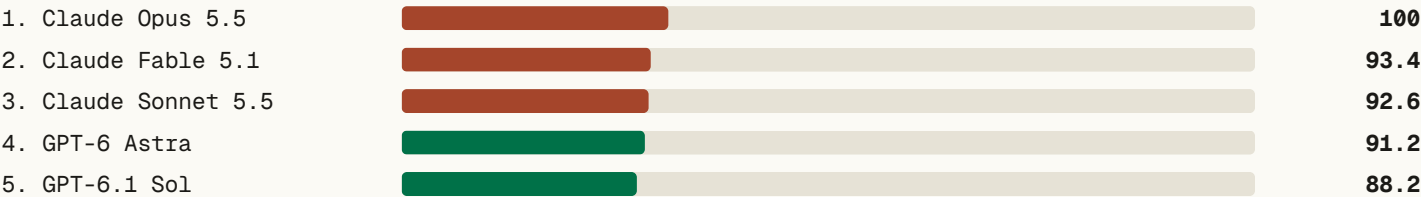
HOW IT COMPARES

MODEL	ROLE	IN / 1M	OUT / 1M	CONTEXT	BENCH INDEX
GLM-5.3 Flash	new	\$0.15	\$0.50	1.05M	52.9
Claude Opus 5.5	today's #1	\$4	\$20	1M	100
Claude Sonnet 5.5	best value	\$2	\$10	1M	92.6
vs Claude Opus 5.5	input 96% cheaper , output 98% cheaper		today's #1: \$4 / \$20, bench index 100; context 1M		
vs Claude Sonnet 5.5	input 93% cheaper , output 95% cheaper		best value: \$2 / \$10, bench index 92.6; context 1M		

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



WHERE THE PUBLIC LEADERBOARDS STAND TODAY



GLM-5.3 Flash sits at 52.9 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 GLM-5.3 Flash as the OpenRouter catalog lists `z-ai/glm-5.3-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.