DeepSeek V4.1 Flash

DeepSeek

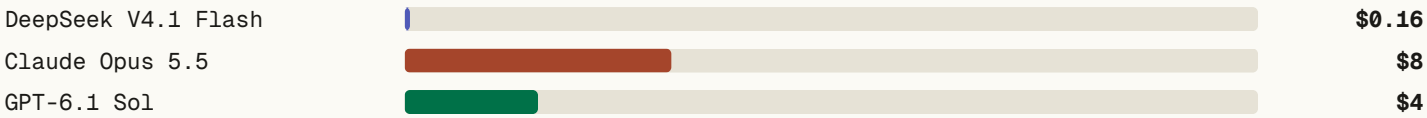
AT A GLANCE

PRICE PER 1M	CONTEXT	MAX OUTPUT	TAKES
$0.045 / $0.50	**1.05M**	**944K**	**text, image**
input / output · cached $0.018	tokens in	tokens per reply	returns text

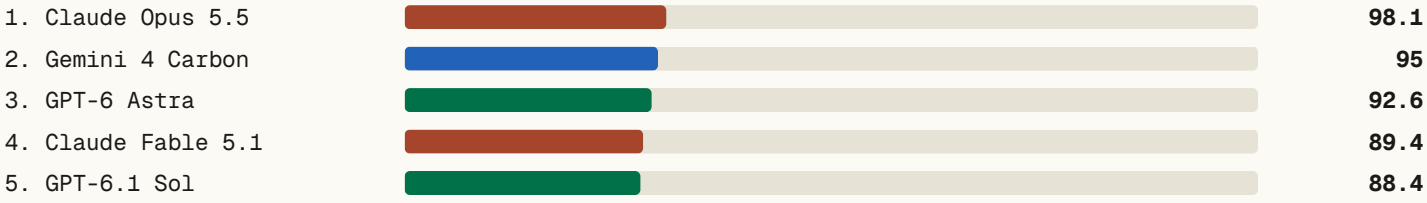
HOW IT COMPARES

MODEL	ROLE	IN / 1M	OUT / 1M	CONTEXT	BENCH INDEX
DeepSeek V4.1 Flash	new	$0.045	$0.50	1.05M	68
Claude Opus 5.5	today's #1	$4	$20	1M	98.1
GPT-6.1 Sol	best value	$2	$10	1.05M	88.4
vs Claude Opus 5.5	**input 99% cheaper, output 98% cheaper**	today's #1: $4 / $20, bench index 98.1; context 1M			
vs GPT-6.1 Sol	**input 98% cheaper, output 95% cheaper**	best value: $2 / $10, bench index 88.4; context 1.05M			

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



WHERE THE PUBLIC LEADERBOARDS STAND TODAY



DeepSeek V4.1 Flash sits at 68 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 DeepSeek V4.1 Flash as the OpenRouter catalog lists deepseek/deepseek-v4.1-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.