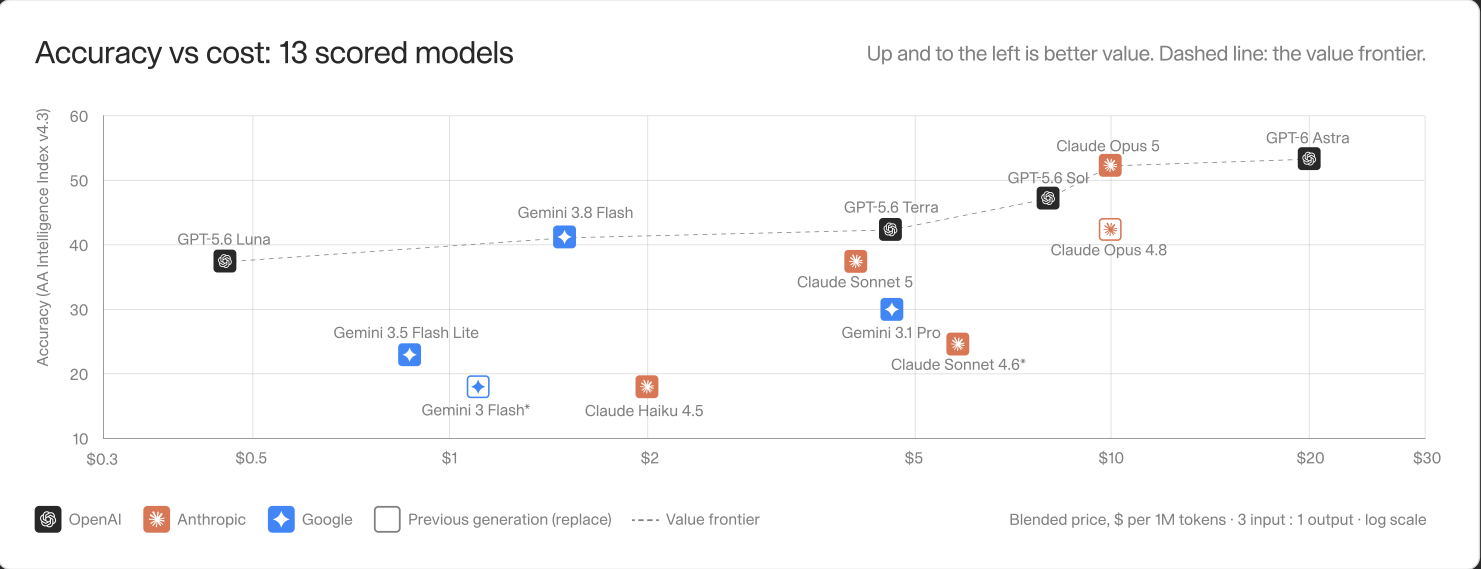


Which LLM for which job

14 models from OpenAI, Anthropic and Google, compared on accuracy, price and accuracy per dollar, with a recommended job for each. Pick the cheapest model that gets the job right, and pay for accuracy only where the step needs it.



How the price works

1x

Input

Instructions, documents, examples, history

~0.1x

Cached input

Repeated prefix read from cache (OpenAI, Anthropic)

5–8x

Output

What the model writes back

=Output

Reasoning

Thinking before the answer, billed as output

BLENDDED PRICE

$(3 \times \text{Input} + \text{Output}) / 4$

GPT-5.6 Sol: $(3 \times \$4 + \$20) / 4 = \$8.00$ per 1M

VALUE

Accuracy / Blended price

GPT-5.6 Luna: $38 / \$0.45 = 84.4$

Multipliers relative to each model's own input price. Batch APIs: 50% off at OpenAI, Anthropic and Google.

Standard API prices · US dollars per 1M tokens								
MODEL	ROLE	RECOMMENDATION	ACCURACY	INPUT	OUTPUT	BLENDED	VALUE	CONTEXT / OUT
OpenAI								
GPT-6 Astra	Frontier	Hardest reasoning, coding and multi-step work. Pay the premium only where the extra accuracy changes the outcome.	53	\$10.00	\$50.00	\$20.00	2.6	1.05M / 128k
GPT-5.6 Sol	Default	Default for demanding work. Near-frontier accuracy at a fraction of the frontier price.	47	\$4.00	\$20.00	\$8.00	5.9	1.05M / 128k
GPT-5.6 Terra	Balanced	Mid-tier option. Luna is cheaper for volume work, Sol more accurate for hard work.	42	\$2.00	\$12.00	\$4.50	9.3	1.05M / 128k
GPT-5.6 Luna	Volume	Best value in the set. Extraction, classification, first-pass review at high volume.	38	\$0.20	\$1.20	\$0.45	84.4	1.05M / 128k
Anthropic								
Claude Opus 5	Writing	Highest-quality writing and analysis. Often verbose, so check cost per job, not per token.	51	\$5.00	\$25.00	\$10.00	5.1	1M / 128k
Claude Opus 5.5	Not yet scored	Newest Opus, not scored in this edition. Lower list price than Opus 5. Test it on your own jobs.	n/s	\$4.00	\$20.00	\$8.00	–	1M / 128k
Claude Opus 4.8	Replace	Previous generation. Claude Opus 5 scores higher at the same list price.	42	\$5.00	\$25.00	\$10.00	4.2	1M / 64k
Claude Sonnet 5	Optional	Strong writing and multi-step work. For pure volume extraction, Luna scores similarly at lower cost.	38	\$2.00	\$10.00	\$4.00	9.5	1M / 128k
Claude Sonnet 4.6	Keep for now	Previous Sonnet. Prefer Claude Sonnet 5 or Luna for new work.	25*	\$3.00	\$15.00	\$6.00	4.2	1M / 64k
Claude Haiku 4.5	Simple tasks	Low-cost Claude for routing and tagging.	18	\$1.00	\$5.00	\$2.00	9.0	200k / 64k
Google								
Gemini 3.8 Flash	Google default	Fast, low cost, strong on long files. Best Google default for volume plus quality.	41	\$0.75	\$3.75	\$1.50	27.3	1.05M / 65.5k
Gemini 3.1 Pro	Long documents	Pro-class model for long documents. 3.8 Flash scores higher at a lower price for most work.	30	\$2.00	\$12.00	\$4.50	6.7	1.05M / 65.5k
Gemini 3.5 Flash Lite	Throughput	Lowest-cost Google option for tagging, routing and short answers.	23	\$0.30	\$2.50	\$0.85	27.1	1.05M / 65.5k
Gemini 3 Flash	Replace	Previous generation. Prefer Gemini 3.8 Flash.	18*	\$0.50	\$3.00	\$1.12	16.0	1.05M / 65.5k

Which model for which job				A starting role for each model. Test on a sample of your own work before assigning it.			
ChatGPT	GPT-5.6 Luna	GPT-5.6 Sol	GPT-6 Astra				
	Volume	Default	Frontier				
	Extraction, classification and first-pass review at scale.	Default for analysis, drafting and questions across several documents.	Use only when the hardest reasoning justifies about 2.5× Sol's blended price.				
Claude	Claude Opus 4.8	Claude Opus 5	Claude Opus 5.5				
	Replace	Writing	Test first				
	Move to Claude Opus 5.	Highest-quality writing and analysis. Measure cost per job: it writes more.	Not scored in this edition. Run it on your own jobs first.				
Gemini	Gemini 3 Flash	Gemini 3.5 Flash Lite	Gemini 3.8 Flash				
	Replace	Throughput	Long documents				
	Move to Gemini 3.8 Flash.	High-throughput tagging, routing and short answers. Test Luna too.	Fast, low-cost work on long files. Watch long-context surcharges elsewhere.				

What one extraction job costs

1,000 fund reports a month · 30,000 input and 2,000 output tokens each. Standard list prices, no caching or batch.

MODEL	COST PER MONTH	ACCURACY
GPT-5.6 Luna	\$8.40	38
Gemini 3.5 Flash Lite	\$14	23
Gemini 3.8 Flash	\$30	41
Claude Sonnet 5	\$80	38
GPT-5.6 Sol	\$160	47
Claude Opus 5	\$200	51
GPT-6 Astra	\$400	53

Add 6,000 reasoning tokens per report and Sol's month rises from \$160 to \$280. Batch halves every figure.

Price watch · review quarterly		
Promotional	GPT-5.6 Sol	\$4/\$20 at least until 21 Nov 2026, then list \$5/\$30. Blended \$8.00 becomes \$11.25.
Introductory	Gemini 3.8 Flash	\$0.75/\$3.75 until 31 Dec 2026, then \$1.50/\$7.50. Blended \$1.50 becomes \$3.00.
Long Inputs	GPT-5.6 models	Over 272k input tokens: 2x input and 1.5x output price.
Long Inputs	Gemini 3.1 Pro	Over 200k tokens: \$4/\$18 instead of \$2/\$12.
Replace	Claude Opus 4.8	Claude Opus 5 scores 9 points higher at the same price.
Replace	Gemini 3 Flash	Gemini 3.8 Flash scores 23 points higher, at \$1.50 vs \$1.12 blended.
Test First	Claude Opus 5.5	Not scored in this edition. \$4/\$20, below Opus 5's list price.

Method. Accuracy = Artificial Analysis Intelligence Index v4.3 score as recorded by V7 on 16 Sep 2026 (* = estimated from a neighbouring index version, n/s = not scored). Blended price = (3 × input + output) / 4. Value = accuracy / blended price, higher is better. Standard API prices in US dollars per 1M tokens from the OpenAI, Anthropic and Google pricing pages, checked 24 Sep 2026. Temporary and long-context prices are listed under Price watch. Scores are general-purpose benchmark results, not accuracy on your documents: test candidate models on a sample of your own work.

V7 Go is the infrastructure layer for AI in document-heavy work: due diligence, underwriting and back-office processes. Model choice is one setting inside its workflows.

