ONTOLOGY EVAL

Which model actually *earns its keep?*

64 AI models measured on the same 37 items of real business and education work — answering from policy documents, building quizzes, writing reports, converting messy PDFs, calling tools and closing the books. Deterministic grading where an answer can be checked by code; a blind four-model panel where it cannot.

CONTENTS

BASIS OF PREPARATION

This report is a print snapshot of the live board at **theontology.ai/eval**, generated from the published v1 results of September 2026. The task set is frozen: new models join the same board as they ship and remain directly comparable with everything already on it. The interactive page carries the per-item detail behind every number printed here.

64	37	9	4
models on the board	items per model, frozen as v1	categories of real work	blind graders per open-ended item

01 Executive summary

GPT-6 Astra leads the board at 0.934 under the published weighting — but the more useful finding is how close, and how workload-dependent, the top of the field now is.

The frontier is crowded. The top five models sit within 0.062 points of each other. At that spacing, which model "wins" depends more on the weighting you apply than on any decisive capability gap — which is why this report scores the same nine category results under four different workload scenarios rather than crowning a single number.

Open weights have closed most of the distance. The best open-weight model, GLM 5.3, scores 0.874 against 0.934 for the best proprietary model — a gap of 0.060 points. That gap is held by a model answered in an agent session rather than through the API (see §10); against the best proprietary model run the same way as the open ones, Claude Fable 5 at 0.881, the gap is 0.007 points. 31 of the 64 models on the board are open weights. One of them has been announced as open but not yet published — Qwen 3.8 Max (due announced for August 2026, not yet on Hugging Face).

Price and quality have come apart. A full run of the eval cost $4.52 on the dearest model and $0.01 on the cheapest — three orders of magnitude — while scores span far less. On cost per point, Gemma 4 31B delivers the board's best value at $0.02 per point of balanced score. For most workloads the economic question is no longer "can we afford the best model?" but "does the work need it?"

Speed is its own axis. Median per-item response times range from 1.1s to 100.9s across the board. The frontier tax is usually paid in seconds as well as dollars.

0.934	0.060	$0.02	1.1s
top score — GPT-6 Astra	gap, best proprietary to best open	best cost per point — Gemma 4 31B	fastest median item — Mercury 2

Scores are on a 0–1 scale: a weighted mean of the model's nine category scores, renormalised over the categories it could attempt. Categories a model cannot attempt (no vision, no function calling) are dropped as n/a rather than counted as failures; refusals and errors score zero. Full method at §10.

02 Leaderboard

Ranked by the balanced weighting — the eval's own published category weights, so this column matches the published composite exactly. Run cost is what the full 37-item run cost at metered prices.

#		MODEL	PROVIDER	WEIGHTS	TIER	RELEASED	BALANCED SCORE	RUN COST	MEDIAN LATENCY	RELIABILITY
1	new	GPT-6 Astra	OpenAI	closed	frontier	Sep 2026	0.934	$3.77	–	100.0%
2	new	Claude Fable 5.1	Anthropic	closed	frontier	Sep 2026	0.913	$4.52	–	100.0%
3	▼2	Claude Fable 5	Anthropic	closed	frontier	Jun 2026	0.881	$4.52	9.1s	100.0%
4	new	GLM 5.3	Z.ai	open	frontier	Aug 2026	0.874	$0.31	15.4s	100.0%
5	▼2	Kimi K3	Moonshot	open	workhorse	Jul 2026	0.872	$1.50	19.3s	100.0%
6	▼2	Claude Opus 5	Anthropic	closed	frontier	Jul 2026	0.871	$2.83	8.5s	100.0%
7	new	Gemini 3.8 Flash	Google	closed	workhorse	Sep 2026	0.867	$0.34	–	100.0%
8	▼6	Kimi K2.6	Moonshot	open	workhorse	Apr 2026	0.867	$0.60	48.3s	100.0%
9	▼2	Claude Opus 4.7	Anthropic	closed	frontier	Apr 2026	0.861	$1.89	4.5s	100.0%
10	▼5	GPT-5.5	OpenAI	closed	frontier	Apr 2026	0.861	$1.39	6.4s	100.0%
11	▼5	Claude Sonnet 5	Anthropic	closed	workhorse	Jun 2026	0.856	$1.15	6.4s	100.0%
12	new	Muse Spark 1.3	Meta	closed	frontier	Sep 2026	0.852	$0.62	10.4s	100.0%
13	▼4	Qwen 3.8 Max	Alibaba	open soon	frontier	Aug 2026	0.849	$1.21	14.6s	100.0%
14	▼6	GPT-5.6 Terra	OpenAI	closed	workhorse	Jul 2026	0.844	$0.43	3.5s	100.0%
15	▼4	Claude Opus 4.8	Anthropic	closed	frontier	May 2026	0.836	$1.93	5.5s	100.0%
16	▼6	GPT-5.6 Sol	OpenAI	closed	frontier	Jul 2026	0.833	$2.29	6.9s	100.0%
17	new	GLM 5.3 Flash	Z.ai	open	budget	Aug 2026	0.831	$0.04	9.3s	100.0%
18	▼6	Qwen 3.7 Max	Alibaba	closed	workhorse	May 2026	0.827	$0.41	14.6s	100.0%
19	new	Grok 4.6	xAI	closed	frontier	Aug 2026	0.821	$0.71	12.0s	100.0%
20	new	Qwen 3.8 Flash	Alibaba	open	budget	Aug 2026	0.817	$0.05	7.7s	100.0%
21		DeepSeek V4 Pro	DeepSeek	open	workhorse	Apr 2026	0.816	$0.19	8.6s	100.0%
22	▼7	Gemini 3.6 Flash	Google	closed	workhorse	Jul 2026	0.815	$0.69	7.2s	100.0%
23	▼10	MiMo v2.5 Pro	Xiaomi	open	budget	Apr 2026	0.815	$0.07	12.1s	100.0%
24	▼8	Gemini 3 Flash	Google	closed	workhorse	Dec 2025	0.813	$0.10	4.0s	100.0%
25	▼8	Grok 4.3	xAI	closed	workhorse	Apr 2026	0.811	$0.26	4.6s	100.0%
26	▼12	GPT-5.6 Luna	OpenAI	closed	budget	Jul 2026	0.810	$0.04	4.9s	100.0%
27		GLM 4.5	Z.ai	open	workhorse	Jul 2025	0.810	$0.25	40.2s	100.0%
28	▼9	GPT-5.4	OpenAI	closed	frontier	Mar 2026	0.809	$0.44	3.3s	100.0%
29	▼11	Kimi K2 Thinking	Moonshot	open	workhorse	Nov 2025	0.804	$0.27	42.5s	100.0%
30	▼6	GLM 5.1	Z.ai	open	workhorse	Apr 2026	0.804	$0.40	20.0s	100.0%
31	▼6	GLM 5.2	Z.ai	open	workhorse	Jun 2026	0.799	$0.10	6.1s	100.0%
32	▼12	Claude Sonnet 4.5	Anthropic	closed	workhorse	Sep 2025	0.794	$1.08	8.9s	100.0%
33	▼7	Grok 4.5	xAI	closed	frontier	Jul 2026	0.794	$0.87	9.9s	100.0%
34	▼12	Kimi K2	Moonshot	open	workhorse	Jul 2025	0.794	$0.08	8.3s	100.0%
35	▼12	GLM 5	Z.ai	open	workhorse	Feb 2026	0.791	$0.27	34.8s	100.0%
36	▼7	Gemini 3.5 Flash Lite	Google	closed	budget	Jul 2026	0.787	$0.08	3.0s	100.0%
37	▼2	LongCat 2.0	Meituan	open	budget	Jul 2026	0.786	$0.13	10.4s	100.0%
38	▼8	Qwen 3.7 Flash	Alibaba	closed	budget	Jul 2026	0.784	$0.02	22.1s	100.0%
39	▼8	Qwen3 Max	Alibaba	closed	workhorse	Sep 2025	0.783	$0.13	6.2s	100.0%
40	▼8	DeepSeek V3.2	DeepSeek	open	workhorse	Dec 2025	0.783	$0.03	9.0s	100.0%
41	▼4	Seed 2.0 Lite	ByteDance	closed	budget	Mar 2026	0.781	$0.19	13.7s	100.0%
42	▼14	MiniMax M3	MiniMax	open	budget	May 2026	0.780	$0.08	8.2s	100.0%
43	▼9	Inkling	Thinking Machines	open	frontier	Jul 2026	0.770	$0.49	11.2s	100.0%
44	▼11	DeepSeek V4 Flash	DeepSeek	open	budget	Apr 2026	0.765	$0.02	8.2s	100.0%
45	▼6	Gemini 3.1 Pro	Google	closed	frontier	Feb 2026	0.760	$1.17	10.7s	100.0%
46	▼10	Claude Sonnet 4.6	Anthropic	closed	workhorse	Feb 2026	0.756	$1.16	7.6s	100.0%
47	▼4	Gemma 4 31B	Google	open	budget	Apr 2026	0.753	$0.01	11.0s	100.0%
48	▼10	DeepSeek R1	DeepSeek	open	workhorse	Jan 2025	0.749	$0.30	100.9s	100.0%
49	▼3	Gemini 2.5 Pro	Google	closed	frontier	Jun 2025	0.742	$1.15	17.6s	100.0%
50	▼8	GPT-5.1	OpenAI	closed	frontier	Nov 2025	0.734	$0.26	4.2s	100.0%
51	▼11	Mistral Large 2512	Mistral	open	workhorse	Dec 2025	0.731	$0.07	5.1s	100.0%
52	▼7	Ling 2.6 1T	InclusionAI	open	workhorse	Apr 2026	0.731	$0.02	2.6s	100.0%
53	▼12	DeepSeek V3 0324	DeepSeek	open	workhorse	Mar 2025	0.728	$0.04	9.5s	100.0%
54	▼10	Claude Sonnet 4	Anthropic	closed	workhorse	May 2025	0.725	$1.06	7.5s	100.0%
55	▼8	Claude Haiku 4.5	Anthropic	closed	budget	Oct 2025	0.714	$0.34	4.4s	100.0%
56	▼8	Nemotron 3 Ultra 550B	NVIDIA	open	workhorse	Jun 2026	0.712	$0.18	5.6s	100.0%
57	▼8	Gemini 2.5 Flash	Google	closed	workhorse	Jun 2025	0.686	$0.09	5.4s	100.0%
58	▼8	Mercury 2	Inception	closed	workhorse	Mar 2026	0.670	$0.04	1.1s	100.0%
59	▼8	ERNIE 4.5 VL 424B	Baidu	open	workhorse	Jun 2025	0.668	$0.10	12.1s	100.0%
60	▼7	GLM 4.7 Flash	Z.ai	open	budget	Jan 2026	0.636	$0.03	16.6s	100.0%
61	▼9	GPT-OSS 120B	OpenAI	open	budget	Aug 2025	0.633	$0.01	8.5s	100.0%
62	▼8	Command A	Cohere	open	workhorse	Mar 2025	0.592	$0.36	7.6s	100.0%
63	▼8	Llama 4 Maverick	Meta	open	budget	Apr 2025	0.578	$0.06	12.8s	100.0%
64	▼8	Llama 4 Scout	Meta	open	budget	Apr 2025	0.484	$0.03	6.8s	100.0%

"new" = added since the board of August 2026; ▲▼ = movement against it. Reliability is the share of items answered without a provider error on the scored attempt; provenance is in Appendix A.3.

03 Scenario weightings

One eval, four weightings. Each scenario reweights the nine categories to match a real workload; nothing is re-run and no new judgement is applied. The weights are published in full below so the arithmetic can be checked — or replaced with your own.

CATEGORY	PER ITEM	BALANCED	EDUCATION	ENGINEERING	BUSINESS OPS
Grounded QA	32%	10%	15%	10%	15%
Quiz generation	5%	11%	25%	–	–
Report writing	3%	10%	10%	–	20%
Summarisation	3%	9%	15%	5%	15%
Retrieval	22%	10%	10%	10%	15%
PDF conversion	11%	7%	10%	5%	15%
Function calling	11%	13%	–	20%	10%
Problem solving	8%	16%	10%	20%	10%
Programming	5%	14%	5%	30%	–

HOW TO READ A SCENARIO SCORE

A scenario score is the weighted mean of a model's nine category scores under that column's weights, renormalised over the categories the model attempted. Nothing is re-run between scenarios; only the weights change. The four leaderboards these weightings produce follow on the next page.

Per item

No editorial weighting — every task carries equal weight.

#	MODEL	SCORE	$/POINT
1	GPT-6 Astra closed	0.968	$3.89
2	Claude Fable 5.1 closed	0.955	$4.74
3	Claude Fable 5 closed	0.944	$4.79
4	Claude Opus 5 closed	0.939	$3.01
5	Kimi K3 open	0.937	$1.60

Balanced

The eval's own published category weighting — the composite.

#	MODEL	SCORE	$/POINT
1	GPT-6 Astra closed	0.934	$4.03
2	Claude Fable 5.1 closed	0.913	$4.95
3	Claude Fable 5 closed	0.881	$5.13
4	GLM 5.3 open	0.874	$0.36
5	Kimi K3 open	0.872	$1.72

Education

Quiz generation and explanation dominate — the classroom workload.

#	MODEL	SCORE	$/POINT
1	Claude Fable 5.1 closed	0.935	$4.83
2	GPT-6 Astra closed	0.923	$4.08
3	Claude Opus 5 closed	0.893	$3.17
4	Kimi K3 open	0.889	$1.69
5	Claude Fable 5 closed	0.889	$5.09

Engineering

Programming against a hidden test suite, tool calling, multi-step problem solving.

#	MODEL	SCORE	$/POINT
1	GPT-6 Astra closed	0.973	$3.87
2	Claude Fable 5.1 closed	0.863	$5.24
3	Claude Fable 5 closed	0.862	$5.24
4	Gemini 3.8 Flash closed	0.860	$0.40
5	GLM 5.3 open	0.859	$0.36

Business ops

Reports, retrieval from policy documents, messy-PDF conversion, closing the books.

#	MODEL	SCORE	$/POINT
1	Claude Fable 5.1 closed	0.970	$4.66
2	Claude Opus 5 closed	0.938	$3.01
3	Kimi K3 open	0.935	$1.61
4	Muse Spark 1.3 closed	0.928	$0.67
5	GPT-5.6 Sol closed	0.918	$2.50

04 The full score matrix

Every category score on the board, sorted by the balanced weighting. Darker is better. Cells marked n/a are categories the model could not attempt — dropped from its composite and reweighted, not counted as failures.

MODEL	GROUNDED QA	QUIZ GENERATION	REPORT WRITING	SUMMARISATION	RETRIEVAL	PDF CONVERSION	FUNCTION CALLING	PROBLEM SOLVING	PROGRAMMING
GPT-6 Astra	1.00	0.92	0.64	0.98	1.00	0.89	1.00	0.97	0.95
Claude Fable 5.1	1.00	0.91	1.00	0.98	1.00	0.89	1.00	0.89	0.64
Claude Fable 5	1.00	0.86	0.80	0.89	1.00	0.88	1.00	0.91	0.64
GLM 5.3	0.92	0.84	0.86	0.90	1.00	n/a	1.00	0.82	0.71
Kimi K3	1.00	0.82	0.87	0.98	1.00	0.88	1.00	0.81	0.60
Claude Opus 5	1.00	0.87	0.93	0.87	1.00	0.88	1.00	0.89	0.50
Gemini 3.8 Flash	1.00	0.78	0.89	0.86	1.00	0.88	0.81	0.85	0.80
Kimi K2.6	1.00	0.78	0.81	0.94	1.00	0.89	1.00	0.80	0.68
Claude Opus 4.7	1.00	0.80	0.90	0.84	1.00	0.88	1.00	0.77	0.66
GPT-5.5	1.00	0.75	0.82	0.97	0.88	0.88	1.00	0.83	0.69
Claude Sonnet 5	1.00	0.82	0.67	0.94	1.00	0.89	1.00	0.86	0.59
Muse Spark 1.3	1.00	0.78	0.84	0.96	1.00	0.88	1.00	0.84	0.50
Qwen 3.8 Max	1.00	0.79	0.65	0.99	1.00	0.89	1.00	0.81	0.62
GPT-5.6 Terra	1.00	0.78	0.85	0.98	0.88	0.88	1.00	0.80	0.55
Claude Opus 4.8	1.00	0.92	0.75	0.90	1.00	0.89	1.00	0.72	0.50
GPT-5.6 Sol	1.00	0.80	0.93	0.98	0.88	0.88	1.00	0.71	0.50
GLM 5.3 Flash	1.00	0.85	0.82	0.91	1.00	0.88	1.00	0.70	0.50
Qwen 3.7 Max	1.00	0.82	0.70	0.99	1.00	n/a	1.00	0.63	0.63
Grok 4.6	1.00	0.78	0.89	0.75	1.00	0.87	1.00	0.74	0.50
Qwen 3.8 Flash	1.00	0.72	0.66	0.93	1.00	0.86	1.00	0.62	0.69
DeepSeek V4 Pro	1.00	0.74	0.86	0.84	1.00	n/a	0.88	0.70	0.64
Gemini 3.6 Flash	1.00	0.84	0.76	0.90	1.00	0.89	1.00	0.65	0.50
MiMo v2.5 Pro	1.00	0.79	0.86	0.91	1.00	n/a	1.00	0.65	0.50
Gemini 3 Flash	1.00	0.76	0.77	0.79	1.00	0.88	1.00	0.75	0.50
Grok 4.3	1.00	0.89	0.71	0.89	1.00	0.88	1.00	0.63	0.50
GPT-5.6 Luna	1.00	0.80	0.84	0.98	0.75	0.86	1.00	0.73	0.49
GLM 4.5	1.00	0.92	0.66	0.92	1.00	n/a	1.00	0.65	0.50
GPT-5.4	1.00	0.88	0.76	0.88	1.00	0.88	0.81	0.75	0.50
Kimi K2 Thinking	1.00	0.77	0.65	0.89	1.00	n/a	1.00	0.73	0.50
GLM 5.1	1.00	0.79	0.80	0.92	1.00	n/a	1.00	0.68	0.42
GLM 5.2	1.00	0.79	0.81	0.82	1.00	n/a	1.00	0.64	0.50
Claude Sonnet 4.5	1.00	0.88	0.78	0.93	1.00	0.86	1.00	0.47	0.50
Grok 4.5	1.00	0.90	0.65	0.69	1.00	0.88	0.88	0.77	0.50
Kimi K2	1.00	0.60	0.82	0.96	1.00	n/a	1.00	0.65	0.50
GLM 5	1.00	0.81	0.66	0.83	1.00	n/a	0.81	0.82	0.50
Gemini 3.5 Flash Lite	0.92	0.73	0.54	0.96	1.00	0.86	1.00	0.72	0.50
LongCat 2.0	1.00	0.90	0.80	0.84	1.00	n/a	0.75	0.68	0.50
Qwen 3.7 Flash	1.00	0.94	0.32	0.92	1.00	0.89	1.00	0.65	0.50
Qwen3 Max	1.00	0.71	0.43	0.87	0.88	n/a	1.00	0.89	0.50
DeepSeek V3.2	1.00	0.66	0.83	0.86	0.88	n/a	1.00	0.68	0.50
Seed 2.0 Lite	1.00	0.96	0.32	0.86	1.00	0.92	1.00	0.63	0.50
MiniMax M3	1.00	0.64	0.63	0.82	1.00	0.89	1.00	0.69	0.50
Inkling	1.00	0.74	0.67	0.78	0.88	0.89	1.00	0.64	0.50
DeepSeek V4 Flash	1.00	0.76	0.67	0.94	1.00	n/a	0.81	0.63	0.50
Gemini 3.1 Pro	1.00	0.78	0.80	0.92	1.00	0.89	1.00	0.31	0.50
Claude Sonnet 4.6	1.00	0.72	0.69	0.98	1.00	0.88	1.00	0.36	0.50
Gemma 4 31B	1.00	0.88	0.35	0.87	0.62	0.88	1.00	0.74	0.50
DeepSeek R1	1.00	0.78	0.63	0.89	1.00	n/a	0.75	0.63	0.50
Gemini 2.5 Pro	1.00	0.94	0.75	0.86	1.00	0.88	0.81	0.30	0.50
GPT-5.1	1.00	0.80	0.64	0.82	0.78	0.86	1.00	0.44	0.50
Mistral Large 2512	1.00	0.68	0.64	0.83	1.00	0.84	0.88	0.76	0.17
Ling 2.6 1T	1.00	0.86	0.40	0.92	0.88	n/a	0.75	0.67	0.50
DeepSeek V3 0324	1.00	0.71	0.54	0.83	1.00	n/a	0.69	0.70	0.50
Claude Sonnet 4	0.92	0.76	0.40	0.88	1.00	0.86	1.00	0.51	0.42
Claude Haiku 4.5	1.00	0.74	0.34	0.91	1.00	0.91	0.88	0.43	0.50
Nemotron 3 Ultra 550B	1.00	0.76	0.63	0.64	1.00	n/a	0.62	0.66	0.50
Gemini 2.5 Flash	1.00	0.70	0.41	0.87	1.00	0.67	0.81	0.42	0.50
Mercury 2	1.00	0.66	0.25	0.53	0.91	n/a	0.88	0.64	0.50
ERNIE 4.5 VL 424B	1.00	0.68	0.31	0.74	0.88	0.89	1.00	0.28	0.50
GLM 4.7 Flash	1.00	0.86	0.40	0.81	0.88	n/a	0.75	0.62	0.00
GPT-OSS 120B	1.00	0.52	0.28	0.54	0.91	n/a	1.00	0.40	0.50
Command A	0.92	0.70	0.38	0.85	0.88	n/a	0.50	0.28	0.50
Llama 4 Maverick	1.00	0.66	0.23	0.75	1.00	0.85	0.31	0.28	0.50
Llama 4 Scout	1.00	0.59	0.18	0.56	0.88	0.87	0.31	0.39	0.00

Shading is scaled to the observed range on this board (0.00–1.00).

05 Category results

Leaders per category, independent of any scenario weighting — the raw material the scenarios are built from. The field average is across every model that attempted the category.

Grounded QA

balanced weight 10%

1	GPT-6 Astra	1.00
2	Grok 4.5	1.00
3	Claude Sonnet 4.5	1.00

field average 0.99

Quiz generation

balanced weight 11%

1	Seed 2.0 Lite	0.96
2	Gemini 2.5 Pro	0.94
3	Qwen 3.7 Flash	0.94

field average 0.79

Report writing

balanced weight 10%

1	Claude Fable 5.1	1.00
2	Claude Opus 5	0.93
3	GPT-5.6 Sol	0.93

field average 0.66

Summarisation

balanced weight 9%

1	Qwen 3.7 Max	0.99
2	Qwen 3.8 Max	0.99
3	Claude Fable 5.1	0.98

field average 0.87

Retrieval

balanced weight 10%

1	GLM 4.5	1.00
2	GPT-5.4	1.00
3	Kimi K2 Thinking	1.00

field average 0.96

PDF conversion

balanced weight 7%

1	Seed 2.0 Lite	0.92
2	Claude Haiku 4.5	0.91
3	GPT-6 Astra	0.89

field average 0.88

Function calling

balanced weight 13%

1	Gemini 3 Flash	1.00
2	Grok 4.3	1.00
3	GPT-5.6 Luna	1.00

field average 0.92

Problem solving

balanced weight 16%

1	GPT-6 Astra	0.97
2	Claude Fable 5	0.91
3	Claude Opus 5	0.89

field average 0.66

Programming

balanced weight 14%

1	GPT-6 Astra	0.95
2	Gemini 3.8 Flash	0.80
3	GLM 5.3	0.71

field average 0.52

open weights · proprietary

06 Cost, value and speed

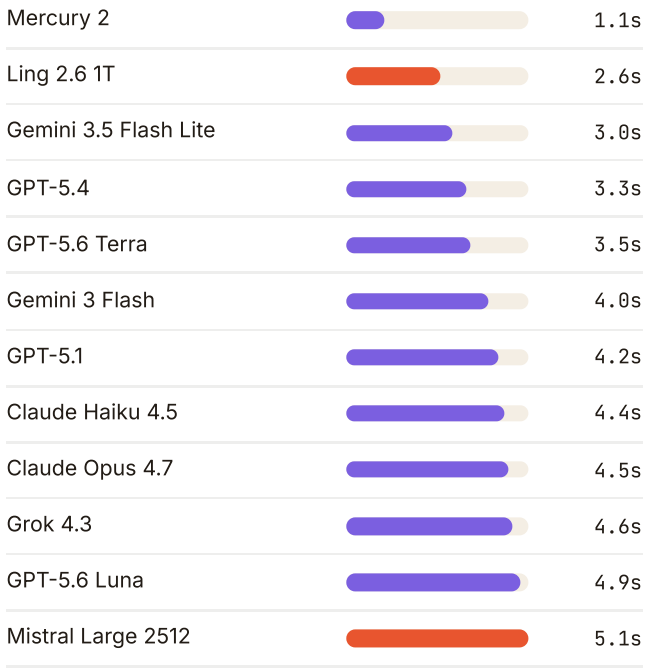
Cost per point of balanced score

Full-run cost divided by balanced score — what a point of quality actually costs to buy. Top twelve of 64.

#	MODEL	$/POINT	RUN COST	SCORE
1	Gemma 4 31B	$0.02	$0.01	0.753
2	GPT-OSS 120B	$0.02	$0.01	0.633
3	Ling 2.6 1T	$0.02	$0.02	0.731
4	DeepSeek V4 Flash	$0.03	$0.02	0.765
5	Qwen 3.7 Flash	$0.03	$0.02	0.784
6	DeepSeek V3.2	$0.03	$0.03	0.783
7	GLM 5.3 Flash	$0.05	$0.04	0.831
8	DeepSeek V3 0324	$0.05	$0.04	0.728
9	GLM 4.7 Flash	$0.05	$0.03	0.636
10	GPT-5.6 Luna	$0.05	$0.04	0.810
11	Qwen 3.8 Flash	$0.06	$0.05	0.817
12	Mercury 2	$0.06	$0.04	0.670

Median latency, fastest twelve

Median per-item response time across the whole run — documents, images and tool calls included, so this is workload latency, not time-to-first-token marketing.



READING THE TWO TOGETHER

The cheapest capable model and the best model are usually not the same model — and the spread between them is wide enough that routing matters: sending routine work to a value model and escalating hard items to a frontier one is, on these numbers, the dominant strategy for any cost-sensitive deployment.

07 Open weights vs proprietary

Group averages are not a fair fight — the two groups differ in size and tier mix. So here is the top of each group under the balanced weighting, and the gap between their best.

Open weights

31 models on the board

#	MODEL	STANDOUT	SCORE
1	GLM 5.3	▲ report writing	0.874
2	Kimi K3	▲ report writing	0.872
3	Kimi K2.6	▲ programming	0.867
4	Qwen 3.8 Max	▲ problem solving	0.849
5	GLM 5.3 Flash	▲ report writing	0.831
6	Qwen 3.8 Flash	▲ programming	0.817
7	DeepSeek V4 Pro	▲ report writing	0.816
8	MiMo v2.5 Pro	▲ report writing	0.815
9	GLM 4.5	▲ quiz generation	0.810
10	Kimi K2 Thinking	▲ function calling	0.804
11	GLM 5.1	▲ report writing	0.804
12	GLM 5.2	▲ report writing	0.799

Best value in group: Gemma 4 31B at $0.02/point.

Proprietary

33 models on the board

#	MODEL	STANDOUT	SCORE
1	GPT-6 Astra	▲ programming	0.934
2	Claude Fable 5.1	▲ report writing	0.913
3	Claude Fable 5	▲ problem solving	0.881
4	Claude Opus 5	▲ report writing	0.871
5	Gemini 3.8 Flash	▲ programming	0.867
6	Claude Opus 4.7	▲ report writing	0.861
7	GPT-5.5	▲ programming	0.861
8	Claude Sonnet 5	▲ problem solving	0.856
9	Muse Spark 1.3	▲ problem solving	0.852
10	GPT-5.6 Terra	▲ report writing	0.844
11	Claude Opus 4.8	▲ quiz generation	0.836
12	GPT-5.6 Sol	▲ report writing	0.833

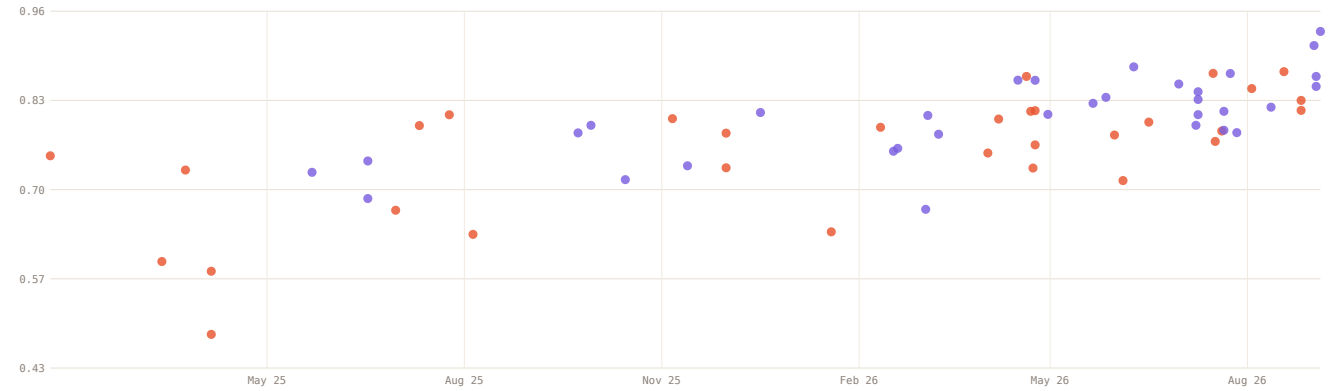
Best value in group: Qwen 3.7 Flash at $0.03/point.

THE GAP

Under the balanced weighting the distance between the best of each group is **0.060 points**, held by GPT-6 Astra. GPT-6 Astra was answered in an agent session, not through the API; against the best proprietary model run the same way as the open ones, Claude Fable 5, the gap is **0.007 points**. "Standout" marks the category where each model sits furthest above the field average — a guide to what each one is *for*, not just where it ranks. For why the open/closed distinction matters in practice, see the companion report, *Where Can You Change an AI?*

08 What a year bought you

Every model on the board carries its release date, so the board can be read as a time series. Each dot is a model under the published weighting. There is no fitted line through the field: a single curve across budget and frontier models at once would describe nothing real. The version-on-version table below is the like-for-like comparison.



open weights proprietary

Version on version

Successive releases of the same model line, which is the only like-for-like comparison over time — a company's releases in date order would run a flagship into a mid-tier model and read as a decline that never happened.

LINE	RELEASES, OLDEST FIRST	FIRST	LATEST	CHANGE
GPT flagship	GPT-5.1 → GPT-5.4 → GPT-5.5 → GPT-5.6 Sol → GPT-6 Astra	0.734	0.934	**+0.199**
GLM Flash	GLM 4.7 Flash → GLM 5.3 Flash	0.636	0.831	**+0.195**
Gemini Flash	Gemini 2.5 Flash → Gemini 3 Flash → Gemini 3.6 Flash → Gemini 3.8 Flash	0.686	0.867	**+0.181**
Claude Sonnet	Claude Sonnet 4 → Claude Sonnet 4.5 → Claude Sonnet 4.6 → Claude Sonnet 5	0.725	0.856	**+0.131**
DeepSeek Chat	DeepSeek V3 0324 → DeepSeek V3.2 → DeepSeek V4 Pro	0.728	0.816	**+0.088**
Kimi	Kimi K2 → Kimi K2 Thinking → Kimi K2.6 → Kimi K3	0.794	0.872	**+0.078**
Qwen Max	Qwen3 Max → Qwen 3.7 Max → Qwen 3.8 Max	0.783	0.849	**+0.066**
GLM	GLM 4.5 → GLM 5 → GLM 5.1 → GLM 5.2 → GLM 5.3	0.810	0.874	**+0.064**
Qwen Flash	Qwen 3.7 Flash → Qwen 3.8 Flash	0.784	0.817	**+0.033**
Claude Fable	Claude Fable 5 → Claude Fable 5.1	0.881	0.913	**+0.032**
Gemini Pro	Gemini 2.5 Pro → Gemini 3.1 Pro	0.742	0.760	**+0.019**
Grok	Grok 4.3 → Grok 4.5 → Grok 4.6	0.811	0.821	**+0.011**
Claude Opus	Claude Opus 4.7 → Claude Opus 4.8 → Claude Opus 5	0.861	0.871	**+0.010**
Llama 4	Llama 4 Maverick → Llama 4 Scout	0.578	0.484	**-0.094**

Read the upward slope with care: two of the nine categories reward knowing what has happened recently, so an older model loses ground on the report task for having an older knowledge cutoff rather than for reasoning worse. Part of the trend is recency. Older models are also measured as their endpoints are served today, which is not necessarily as they launched.

09 Checking the graders for bias

The five open-ended items are about a third of the composite and the part a model could be flattered on, so four models mark them blind under shuffled labels, and all four have a model on the board. Mapping the labels back shows how each grader rated each model — on the ringed diagonal, its own. This is an audit of the marking, not a ranking: rows are in board order, and a model that leads on the checkable categories can sit well down the rubric means.

MODEL	PANEL	CLAUDE-FABLE-5 +0.02	CLAUDE-OPUS-5 +0.01	GPT-5.6-SOL +0.04	QWEN-3.8-MAX -0.07
🇺🇸 GPT-6 Astra	**4.24**	4.24	4.24	4.48	4.00
🇺🇸 Claude Fable 5.1	**4.55**	4.48	4.62	4.57	4.52
🇺🇸 Claude Fable 5	**4.86**	**4.86**	4.81	4.81	4.95
🇨🇳 GLM 5.3	**4.35**	4.29	4.38	4.43	4.29
🇨🇳 Kimi K3	**4.71**	4.76	4.81	4.48	4.81
🇺🇸 Claude Opus 5	**4.86**	4.95	**4.90**	4.86	4.71
🇺🇸 Gemini 3.8 Flash	**4.02**	4.05	4.05	4.10	3.90
🇨🇳 Kimi K2.6	**4.39**	4.48	4.38	4.57	4.14
🇺🇸 Claude Opus 4.7	**4.43**	4.67	4.38	4.29	4.38
🇺🇸 GPT-5.5	**4.18**	4.10	3.95	4.29	4.38
🇺🇸 Claude Sonnet 5	**4.51**	4.38	4.38	4.81	4.48
🇺🇸 Muse Spark 1.3	**4.35**	4.19	4.48	4.43	4.29
🇨🇳 Qwen 3.8 Max	**4.21**	4.29	4.19	4.10	**4.29**
🇺🇸 GPT-5.6 Terra	**4.40**	4.52	4.29	4.38	4.43
🇺🇸 Claude Opus 4.8	**4.49**	4.48	4.48	4.48	4.52
🇺🇸 GPT-5.6 Sol	**4.56**	4.67	4.52	**4.62**	4.43
🇨🇳 GLM 5.3 Flash	**4.76**	4.76	4.81	4.76	4.71
🇨🇳 Qwen 3.7 Max	**4.45**	4.48	4.43	4.43	4.48
🇺🇸 Grok 4.6	**3.98**	4.00	3.90	3.76	4.24
🇨🇳 Qwen 3.8 Flash	**3.49**	3.43	3.48	3.67	3.38
🇨🇳 DeepSeek V4 Pro	**3.82**	4.24	4.00	3.71	3.33
🇺🇸 Gemini 3.6 Flash	**4.25**	4.10	4.19	4.33	4.38
🇨🇳 MiMo v2.5 Pro	**4.23**	4.48	4.33	4.19	3.90
🇺🇸 Gemini 3 Flash	**3.77**	3.86	3.81	3.62	3.81
🇺🇸 Grok 4.3	**3.83**	3.86	4.00	3.71	3.76
🇺🇸 GPT-5.6 Luna	**4.44**	4.52	4.43	4.43	4.38
🇨🇳 GLM 4.5	**3.92**	3.86	4.14	3.86	3.81
🇺🇸 GPT-5.4	**4.00**	4.14	3.95	3.90	4.00
🇨🇳 Kimi K2 Thinking	**3.65**	3.67	3.86	3.62	3.48
🇨🇳 GLM 5.1	**4.11**	4.19	4.05	4.10	4.10
🇨🇳 GLM 5.2	**4.07**	4.05	4.00	4.10	4.14
🇺🇸 Claude Sonnet 4.5	**3.74**	3.71	3.76	3.90	3.57
🇺🇸 Grok 4.5	**3.64**	3.67	3.76	3.67	3.48
🇨🇳 Kimi K2	**3.79**	3.76	3.81	3.90	3.67
🇨🇳 GLM 5	**4.00**	3.90	4.05	4.14	3.90
🇺🇸 Gemini 3.5 Flash Lite	**3.80**	3.86	3.81	3.90	3.62
🇨🇳 LongCat 2.0	**3.99**	4.05	3.81	4.19	3.90
🇨🇳 Qwen 3.7 Flash	**3.80**	3.71	3.81	3.81	3.86
🇨🇳 Qwen3 Max	**3.42**	3.38	3.33	3.62	3.33
🇨🇳 DeepSeek V3.2	**3.36**	3.38	3.43	3.48	3.14
🇨🇳 Seed 2.0 Lite	**3.79**	3.81	3.86	3.86	3.62
🇨🇳 MiniMax M3	**3.69**	3.86	3.71	3.81	3.38
🇺🇸 Inkling	**3.77**	3.95	3.76	3.81	3.57
🇨🇳 DeepSeek V4 Flash	**3.86**	4.05	3.90	3.52	3.95
🇺🇸 Gemini 3.1 Pro	**4.20**	4.19	4.05	4.33	4.24
🇺🇸 Claude Sonnet 4.6	**4.27**	4.24	4.48	4.14	4.24
🇺🇸 Gemma 4 31B	**3.67**	3.71	3.62	3.67	3.67
🇨🇳 DeepSeek R1	**3.88**	3.71	3.86	3.86	4.10
🇺🇸 Gemini 2.5 Pro	**4.24**	4.33	4.19	4.19	4.24
🇺🇸 GPT-5.1	**4.02**	4.14	3.95	3.95	4.05
🇫🇷 Mistral Large 2512	**3.37**	3.48	3.33	3.43	3.24
🇨🇳 Ling 2.6 1T	**3.88**	4.00	4.05	3.90	3.57
🇨🇳 DeepSeek V3 0324	**3.30**	3.24	3.24	3.43	3.29
🇺🇸 Claude Sonnet 4	**3.62**	3.57	3.71	3.52	3.67
🇺🇸 Claude Haiku 4.5	**3.42**	3.57	3.43	3.38	3.29
🇺🇸 Nemotron 3 Ultra 550B	**3.39**	3.62	3.52	3.43	3.00
🇺🇸 Gemini 2.5 Flash	**3.33**	3.19	3.29	3.67	3.19
🇺🇸 Mercury 2	**2.76**	3.00	2.76	2.76	2.52
🇨🇳 ERNIE 4.5 VL 424B	**3.02**	2.95	2.90	3.19	3.05
🇨🇳 GLM 4.7 Flash	**3.19**	3.14	3.14	3.24	3.24
🇺🇸 GPT-OSS 120B	**2.13**	2.14	2.19	2.43	1.76
🇨🇦 Command A	**3.18**	3.10	3.19	3.29	3.14
🇺🇸 Llama 4 Maverick	**2.62**	2.48	2.38	3.05	2.57
🇺🇸 Llama 4 Scout	**2.05**	1.71	2.05	2.52	1.90

Scores are means over every rubric dimension, 0–5. The figure under each grader is how far that grader sits from the panel overall — a strict grader marks everything low, which is different from disliking a particular model. Ringed cells are a grader marking its own model: claude-fable-5 +0.00 · claude-opus-5 +0.05 · gpt-5.6-sol +0.06 · qwen-3.8-max +0.07.

10 Methodology

Versioned, not rounds

The 37-item task set is frozen as v1. New models are added to this board as they ship and stay directly comparable with everything already on it. When the tasks themselves change, that becomes v2 — a fresh board, not a reshuffle of this one.

Nine categories, five weightings

Every model gets one score per category; the scenarios are different weightings over the same nine numbers. **Per item** is the derived reference — no editorial judgement, every task counting equally, so a category is worth exactly as many items as it holds.

The published weighting follows discrimination, and changed on 5 August 2026

Weight now sits where models actually differ. Grounded QA has twelve items but 53 of 56 models score a perfect 1.00 on it — a gate everyone passes. Programming, problem solving and report writing spread the field by 0.63 to 0.75 and were underweighted relative to what they reveal. QA fell from 18% to 10% and retrieval from 14% to 10%; problem solving rose to 16% and programming to 14%. The spread across the field widened from 0.33 to 0.40. The change moved 46 of 56 models and moved Claude Fable 5 from third to first — said plainly, because a weighting revised after seeing results, in a direction that restores a previous leader, is exactly what a reader should be suspicious of. The rule was fixed before the board was recomputed and the per-item view is published alongside. It also made the eval less precise per item: dropping any single item now moves a score by ±0.019, not ±0.014.

Deterministic first

Everything numeric, exact-match, table F1, the Ruby test suite and forecast RMSE are graded by code. Confabulation traps — questions the source cannot answer — are scored on whether the model says so. Code has format assumptions and they get found: on 5 September 2026 the quiz checker was found to count an answer key's numbered lines as extra questions, to miss options written as Markdown bullets, and to miss explanations given inline after each keyed answer. Fixed, it raised quiz scores on 46 items across 41 models and moved 39 composites, all upward and none by more than 0.037; the graders' rubric marks were unaffected.

Blind where it matters

Open-ended items are reviewed under shuffled anonymous labels by graders who have never seen the answer key, scored 0–5 on fixed rubric dimensions, and only unblinded after scores are locked.

Four graders, not one

Every open-ended item is scored independently by four models — Claude Fable 5, Claude Opus 5, GPT-5.6 Sol and Qwen 3.8 Max — each on the same shuffled pack; the published rubric score is the mean of the four. On the main pack every pair of graders agrees between 0.71 and 0.88; on the September pack, between 0.75 and 0.89. The two Claude graders are closest both times. Adding the fourth grader moved seventeen of the fifty-six models then on the board and changed the average score by 0.001.

Eight models added in September 2026, graded against anchors

The eight additions were reviewed as their own blind pack: their answers plus ten anchor candidates the panel had already scored in August, spanning the board from Llama 4 Scout to Claude Opus 5, reshuffled under fresh labels. Comparing each grader's two scores for the same anchor text measures the drift between packs: the panel marked the September pack 0.22 points stricter on the 0–5 scale (Fable 5 by 0.09, Sol 0.17, Opus 0.27, Qwen 0.35) while agreeing with its August self at r = 0.82–0.88. The new models are published as marked. Adding the shift back would raise their composites by 0.003 to 0.008 — inside the ±0.019 resolution — moving Gemini 3.8 Flash up two places and Grok 4.6, GLM 5.3 Flash and Muse Spark 1.3 up one, with nothing changing at the top. The anchors' August scores stand.

The graders were audited for self-interest

All four graders have a model on this board, so grader-vs-self bias was measured directly. Across the twenty rubric scores each gives itself, the gap averages +0.04 on a 0–5 scale: Opus +0.05, Sol +0.06, Qwen +0.07, and Claude Fable 5 scores itself exactly what the others do. Family is the sharper test now that the graders' successors are on the board: GPT-5.6 Sol marks GPT-6 Astra +0.24 above the panel, the largest lean in the matrix, while Qwen marks the same answers −0.24 below it; the Claude graders show no such lean toward Claude Fable 5.1 (Fable 5 −0.07, Opus +0.07). Qwen 3.8 Max grades through the API with no access to the repository, so its blinding is structural rather than promised; at fifty-six candidates it could not hold the quiz pack in a single pass, so that item was graded in groups of eight against the same rubric — a narrower comparison field than the other three had. Averaging four graders keeps any one leaning to a quarter of its weight.

When a model said nothing, we found out why

Three items came back empty: the request succeeded, the whole completion budget went on hidden reasoning, and no answer appeared. Raising the budget did not help — it bought more thinking. Capping the *reasoning* instead, leaving room to reply, produced an answer on all three within a few thousand tokens. The September additions needed the same treatment on six items across GLM 5.3, GLM 5.3 Flash and Qwen 3.8 Flash; two routes ignored an effort level and honoured only a token budget. GLM 5.3 on the inline forecast item returned nothing at 32,000 tokens on four attempts across three providers that all ignored the cap — reasoning cannot be switched off on that model — and answered in under 5,000 tokens on a fourth provider that honours it. Those caps are recorded per model in the registry and applied to every item that model runs. A reasoning model can be unable to answer at any budget and answer easily at a smaller one.

Three models were answered in agent sessions

Claude Fable 5.1, GPT-6 Astra and Gemini 3.8 Flash were run as Claude Code, Codex and Antigravity sessions from a pack exported by the harness — the same system prompt, user message and attachments as every API run, under rules forbidding web search, code execution and any file outside the pack, with the tool-use items answered through a call shim exposing the same four functions. What differs is the wrapper: an agent harness rather than a bare API call, no metered usage, no timing. Their run cost is taken from the tokens each model's own predecessor spent per item through the API, priced at the new model's list price, and they sit out of the latency figures.

Honest n/a

Text-only models are not punished for vision items, nor endpoints without function calling for tool items. Those categories are dropped and the composite reweighted — refusals and errors, though, score zero.

Re-runs, and why

The first pass had two faults of the harness's own making: a token cap set too low let reasoning models spend their whole budget thinking and return nothing visible, and a provider guardrail truncated some responses mid-sentence. Both are properties of the harness and the route, not the models, so affected items were re-run with a higher cap. Every re-run item is counted; the raw first-pass results are kept.

Where a route failed entirely

A handful of items kept returning truncated output on every provider serving that model. Rather than score a model zero for its router, those answers were produced through a direct session. They are not like-for-like with the provider runs — different plumbing, no document-attachment path — and the models carrying them are listed in Appendix A.3.

A Appendix

A.1 Category definitions

High-school question answering
Grounded QA · weight 10.0%

Twelve freshly authored questions across maths, physics, chemistry and biology, pitched at Australian Years 10-12. Eight are pure computation; four require interpreting supplied data. Graded on the final answer only.

Quiz creation
Quiz generation · weight 11.0%

Two assessment-authoring briefs. Scored half on machine-checkable structure (item counts, answer-key validity, correctness of keyed answers) and half on blind review of pedagogical quality.

Sector report writing
Report writing · weight 10.0%

One briefing on the Australian higher education sector. Scored on blind rubric plus a factual-claims probe — invented statistics are penalised.

Document summarisation
Summarisation · weight 9.0%

Summarise an 11-page public report. Scored on blind rubric plus recall probes for facts that must survive compression, and a faithfulness check for claims the document does not make.

Key-information retrieval
Retrieval · weight 10.0%

Eight questions against a 12-page public document: three direct lookups, three requiring combination across the document, two whose answers are not in the document at all.

Document to Markdown
PDF conversion · weight 7.0%

Four pages converted to Markdown — prose, a dense table, a two-column layout, and a degraded scan. Text-only models score n/a and their composite is reweighted across the remaining categories.

Tool use
Function calling · weight 13.0%

Four business questions answered against a fictional company's data via four tools. Includes a no-tools control and a question the data cannot answer.

Business problem solving
Problem solving · weight 16.0%

Optimisation, forecasting and break-even scenarios with verifiable numeric answers.

Programming
Programming · weight 14.0%

A Ruby class graded by a hidden test suite, and an R statistical model graded by out-of-sample forecast error. Both are executed, not read.

A.2 Score construction

A model's category score is the mean of its item scores in that category, each on a 0–1 scale. Its scenario score is the weighted mean of category scores under the scenario's weights, renormalised over the categories the model attempted. Cost per point divides full-run cost by scenario score. All prices are metered usage at the provider's published rates at run time.

A.3 Provenance — re-runs, direct items and session runs

Models whose published score depended on a second attempt after a harness or provider fault, on items answered outside the provider path, or on a whole run answered in an agent session rather than through the API. Session runs saw the same prompts and attachments as every API run; their cost is taken from their predecessor's token use at list price, and their latency was not measured.

MODEL	RE-RUN	DIRECT	RUN THROUGH
GPT-6 Astra	–	–	Codex session
Claude Fable 5.1	–	–	Claude Code session
Gemini 3.8 Flash	–	–	Antigravity session
Grok 4.6	13	–	API
Claude Fable 5	2	4	API
MiniMax M3	3	–	API
Qwen 3.8 Flash	3	–	API
GLM 5.3	3	–	API
GLM 5.3 Flash	2	–	API
Kimi K3	2	–	API

MODEL	RE-RUN	DIRECT	RUN THROUGH
MiMo v2.5 Pro	2	–	API
Muse Spark 1.3	1	–	API
Claude Opus 5	1	–	API
DeepSeek V4 Pro	1	–	API
Kimi K2 Thinking	1	–	API
LongCat 2.0	1	–	API
Inkling	1	–	API
DeepSeek V4 Flash	1	–	API
GLM 4.7 Flash	1	–	API
GPT-OSS 120B	1	–	API

A.4 Use of this report

This report is published free to read and free to quote, with attribution to theontology.ai/eval. Scores measure performance on this task set under this method only; they are not a general fitness ranking, a safety assessment, or advice for any particular procurement decision. Model names and prices are as published by their providers at the time of the run and may have changed since. Corrections and arguments are welcome: hello@theontology.ai.