Sample report: six example tasks, ranked by quality, cost and speed

6 tasks × 3 models × 5 runs, graded blind. A run is successful when every hard rule passes and it meets at least 80% of the checks.

Recommendation

Best quality: a tie between GPT-6 Astra and Gemini 4 Carbon, which met 97.3% and 96.3% of your checks on average; the gap is within run-to-run noise. GPT-6 Astra succeeded on 30 of 30 runs (100%).

Best value: also GPT-6 Astra, at $0.0078 per successful task, counting the cost of failed attempts.

Fastest: GPT-6 Astra, with a median of 3.6 s per answer (9 in 10 answers within 4.8 s).

Claude Opus 5.5 succeeded on too few runs to be considered for value or speed (under 80% of the best pass rate).

Recommendation: GPT-6 Astra. It is at the top on quality (tied) and costs least per successful task, and it is also the fastest.

This applies to these tasks only. It is not a general ranking of the models.

Ranking

#	Model	Quality	Successful	Cost / successful	Cost / attempt	Latency med / p90	Consistent
1	GPT-6 Astra	97.3% (80-100)	30/30	$0.0078	$0.0078	3.6 s / 4.8 s	6/6
2	Gemini 4 Carbon	96.3% (80-100)	30/30	$0.0083	$0.0083	15 s / 27 s	6/6
3	Claude Opus 5.5	92.6% (50-100)	23/30	$0.0092	$0.0071	3.9 s / 5.2 s	3/6

Task by task

1. Support reply: late refund

Model	Run scores	Successful	Mean	Cost / successful	Median latency
GPT-6 Astra	90 · 90 · 80 · 80 · 80	5/5	84%	$0.0097	4.5 s
Gemini 4 Carbon	80 · 80 · 80 · 80 · 80	5/5	80%	$0.012	18 s
Claude Opus 5.5	70 · 80 · 80 · 70 · 70	2/5	74%	$0.028	5.3 s

2. SQL: revenue per customer

Model	Run scores	Successful	Mean	Cost / successful	Median latency
GPT-6 Astra	100 · 100 · 100 · 100 · 100	5/5	100%	$0.013	4.8 s
Gemini 4 Carbon	88.9 · 100 · 100 · 100 · 100	5/5	97.8%	$0.0084	23 s
Claude Opus 5.5	100 · 100 · 100 · 100 · 100	5/5	100%	$0.0089	3.8 s

3. Summary: incident report

Model	Run scores	Successful	Mean	Cost / successful	Median latency
GPT-6 Astra	100 · 100 · 100 · 100 · 100	5/5	100%	$0.012	3.6 s
Gemini 4 Carbon	100 · 100 · 100 · 100 · 100	5/5	100%	$0.018	24 s
Claude Opus 5.5	85.7 · 100 · 100 · 85.7 · 85.7	2/5	91.4%	$0.029	4.7 s

4. Classify: support ticket

Model	Run scores	Successful	Mean	Cost / successful	Median latency
GPT-6 Astra	100 · 100 · 100 · 100 · 100	5/5	100%	$0.0010	1.7 s
Gemini 4 Carbon	100 · 100 · 100 · 100 · 100	5/5	100%	$0.0018	7.0 s
Claude Opus 5.5	100 · 100 · 100 · 100 · 100	5/5	100%	$0.0006	1.4 s

5. Extract: purchase order to JSON

Model	Run scores	Successful	Mean	Cost / successful	Median latency
GPT-6 Astra	100 · 100 · 100 · 100 · 100	5/5	100%	$0.0054	1.9 s
Gemini 4 Carbon	100 · 100 · 100 · 100 · 100	5/5	100%	$0.0071	12 s
Claude Opus 5.5	100 · 100 · 100 · 100 · 100	5/5	100%	$0.0037	1.7 s

6. Pushback (beta): missing information (beta: pushback)

Model	Run scores	Successful	Mean	Cost / successful	Median latency
GPT-6 Astra	100 · 100 · 100 · 100 · 100	5/5	100%	$0.0052	3.4 s
Gemini 4 Carbon	100 · 100 · 100 · 100 · 100	5/5	100%	$0.0024	7.9 s
Claude Opus 5.5	100 · 100 · 50 · 100 · 100	4/5	90%	$0.0081	4.1 s

Method

Same messages for every model through the Codegraff gateway (The Standard Harness Company of Singapore); 5 runs per task and model; failed calls retried up to 3 times and every attempt's cost counted. Rule checks by code; checklist graded blind by GPT-6 Sol (gpt-6-sol), never one of the models compared. Cost per successful task = all spend on attempts / successful attempts. Prices: list price per token from the gateway. Grading cost not included. Synthetic example tasks; no customer data.