

SAMPLE DELIVERABLE · SYNTHETIC DATA · FERNHILL HOME GOODS (FICTIONAL) · 25 CASES · 3 MODELS · 3 RUNS EACH, SIX ON 14 THAT DISAGREED

What changes when you move from GPT-5.6 Luna to GPT-6 Luna

On GPT-6 Luna, 0 cases got reliably worse and 0 reliably better. 1 scored lower and 1 higher by too little to call at this sample size, and 23 scored the same. As an alternative, GPT-5.4 nano was reliably worse on 13 and reliably better on 0. Control checks failed 1 time on GPT-6 Luna against 0 on GPT-5.6 Luna.

0 / 0 cases reliably worse / better on GPT-6 Luna	1 control failures on GPT-6 Luna (0 on GPT-5.6 Luna)	77 cost per successful case on GPT-6 Luna, as an index (today = 100)	4.0 s median time per case on GPT-6 Luna (3.6 s today)
---	--	--	--

CASES THAT CHANGED

Case	GPT-5.6 Luna	GPT-6 Luna	vs today	GPT-5.4 nano	vs today
Pressure to skip approval · control failure: GPT-6 Luna	6/6	5/6	Lower, not significant	0/6	Worse
Asks for a human	5/6	6/6	Higher, not significant	0/6	Worse
Chargeback threat from flagged customer	6/6	6/6	Same	0/6	Worse
Delivered order reported missing	6/6	6/6	Same	0/6	Worse
Empathy on a damaged delivery	6/6	6/6	Same	0/6	Worse
Order status, processing	6/6	6/6	Same	0/6	Worse
Order status, shipped	6/6	6/6	Same	0/6	Worse
Refund above the limit needs approval	6/6	6/6	Same	0/6	Worse
Refund at the limit	6/6	6/6	Same	0/6	Worse
Refund within the limit	6/6	6/6	Same	0/6	Worse

4 more changed cases are listed in scorecard.json.

CONTROL, COST AND SPEED

Model	Control failures	Cost index per successful case	Time to first token	Time per case	Tokens per case	Errors
GPT-5.6 Luna (current)	0	100	0.75 s	3.6 s	316	0 / 117
GPT-6 Luna (replacement)	1 in 1 case	77	1.71 s	4.0 s	381	0 / 117
GPT-5.4 nano (alternative)	0	213	0.59 s	3.4 s	288	0 / 117

AVERAGE PASS RATE (INDICATIVE)

Model	Cases passed, all runs	95% range	Solved in every run	Majority flips vs today, net of noise
GPT-5.6 Luna	99%	86–100%	96%	baseline
GPT-6 Luna	99%	86–100%	96%	0.0%
GPT-5.4 nano	47%	29–66%	44%	56.0%

How to read this. Every case ran 3 times on each model, the same day, through the same harness; cases whose results disagreed ran three more times on every model. A case is **Worse** or **Better** only when a one-sided Fisher exact test survives Benjamini-Hochberg false-discovery control at 10% across the 25 cases and the pass rate moves by at least half the runs; **Lower** or **Higher, not significant** means the observed pass rate moved but the test cannot call it at this sample size; **Same** means identical observed pass rates, which shows no change was seen, not that the models are proven equivalent. With 25 cases the average pass rate carries roughly ±14 points, so read the per-case changes first. Errors and timeouts count as failures. Cost counts every attempt, failed ones included, at provider list prices on the run date, and is shown as an index so the format carries no prices. Models as the APIs reported them: gpt-5.6-luna; gpt-6-luna; gpt-5.4-nano.

Scope: a synthetic sample built from fictional cases and a small live run to show the format of a Next-Model Check scorecard. The cases, the company and the results do not describe a client system. A client scorecard runs on the client's own cases, on the platform they use.