Four-model cost comparison

19 September 2026 | Trivia delegation experiment | Medium reasoning

All four scored 10/10. Luna cost the least.

Twelve fresh agents answered the same ten fixed trivia questions: three agents per model, each replying once. All answers were correct, with no corrections or retries. Costs below are estimates from logged agent tokens and standard API rates, not actual Codex subscription charges.

Model	Total tokens	Estimated USD	Relative cost
Luna	148,345	$0.008643	1.00x
Terra	107,072	$0.055758	6.45x
Sol	107,067	$0.111737	12.93x
Astra	94,301	$0.140242	16.23x

Where the tokens went

Model	Uncached input	Cached input	Output	Reasoning*
Luna	29,206	118,784	355	249
Terra	18,693	88,320	59	0
Sol	18,815	88,192	60	0
Astra	4,767	89,472	62	0

* Reasoning is included in output, not added again. Total input equals uncached plus cached input. The other models logged zero reasoning tokens despite medium configuration; medium is a setting, not a fixed token budget.

What this tells us

Luna used the most tokens but had the lowest estimated cost. Terra cost 6.45 times Luna, Sol 12.93 times, and Astra 16.23 times in this run. Simple trivia did not reveal a quality advantage for the larger models; it does not establish how they perform on difficult, long-running work.

Comparison limits: Full input sizes and cache coverage differed, especially favoring Astra. These are run-specific results. Coordinator and voice usage are excluded.

Method and scope

Corresponding agents received identical task prompts with no conversation-history fork and no tools. Ten questions were split into groups of four, three, and three, covering astronomy, biology, chemistry, botany, geography, geometry, grammar, arithmetic, and music. Model and medium reasoning settings were verified in local session logs.

This is a new matched-prompt trial. Earlier exploratory trials used eight agent turns per model, generated questions rather than answering fixed questions, and used low reasoning for Luna. Those earlier runs are excluded.

Price assumptions

Per million tokens	Luna	Terra	Sol	Astra
Uncached input	$0.20	$2.00	$4.00	$10.00
Cached input	$0.02	$0.20	$0.40	$1.00
Output	$1.20	$12.00	$20.00	$50.00

Estimated cost = (uncached input x input rate + cached input x cache rate + output x output rate) / 1,000,000. Cache-write counts were zero. Standard short-context API pricing is assumed, without regional surcharges or fast-mode premiums. Every request was below the long-context threshold.

Sensitivity to caching

Pricing the same measured inputs entirely as uncached gives the hypothetical amounts below. This is not a second measurement and still does not normalize the differing input sizes.

Model	Luna	Terra	Sol	Astra
No-cache estimate	$0.030024	$0.214734	$0.429228	$0.945490

Identical task prompts did not produce identical full input token counts. The cause was not isolated. A stronger benchmark would use representative real tasks, repeated trials, controlled context and caching, and include coordinator overhead. These estimates do not measure Codex subscription allowance deductions.

Sources and supporting data

Official OpenAI pricing: Luna | Terra | Sol | Astra. Rates checked 19 September 2026.

Measured counters: companion file model-comparison-data.json. This PDF presents the findings from four-model-cost-report.md.