GPT-4o mini

Performance on the CRED single-error benchmark

Overall performance

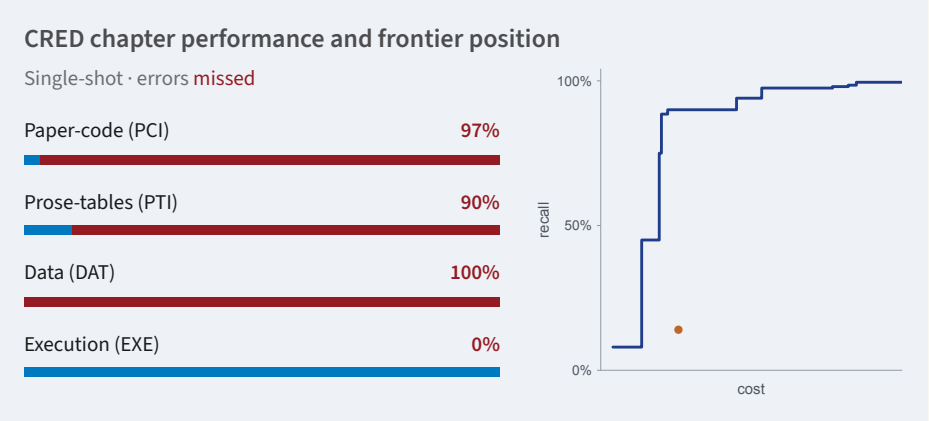
GPT-4o mini detects just 14% of the injected errors. Its low overall recall is not confined to a difficult corner of the benchmark: it misses most paper-code and prose-table inconsistencies, and every injected data-integrity defect. Successful execution checks account for most of its credited detections.

This assessment uses the CRED taxonomy and controlled error-injection framework introduced in *Verifying the Verifiers*. The model receives the manuscript and code in one request through OpenRouter, together with a supplied code-running verdict. There are 100 host papers, each prepared in two separately evaluated single-error versions, giving 200 trials.

The tested configuration has no reasoning-effort ladder. Its reconstructed cost is $0.004 per paper: inexpensive per call, but not competitive verification. The comparison uses the frontier dated 06-SEP-2026, rather than the alternatives available when this older model was released.

gpt-oss-120b detects 45% at a lower cost. GLM 5.3 Flash xhigh reaches 90% for about $0.003 per paper, also less than GPT-4o mini. A small request bill therefore does not establish cost-effectiveness: the benchmark contains configurations that catch far more at still lower cost.

The limited recall cannot simply be attributed to a paper-only test: this configuration receives code too. Nor are its misses confined to relationships that require tracing an analysis script. It struggles even with prose-table contradictions, where the necessary comparison lies within the written results.



GPT-4o mini misses most errors outside the supplied execution checks; the frontier offers much higher recall at lower cost.

Where it is weak

The model detects only 3 of 90 paper-code defects and 5 of 50 prose-table defects. It detects none of the 40 data-integrity defects. These results point to broad difficulty checking the supplied artifacts, rather than a single narrow weakness in variable construction or statistical interpretation.

All 20 execution defects receive detection credit. But those tests supply a code-running verdict; the model is not independently reproducing the research. Remove that chapter and it detects only eight of 180 injected errors. The headline average conceals how little detection remains across the other three chapters.

The example illustrates another risk: producing plausible-sounding objections without catching the actual contradiction. This is a qualitative observation from an inspected response, not an estimated false-positive rate. The benchmark scores recall; an audit of all reported objections would be needed to measure precision.

A symptomatic miss

In a benchmark copy of a study of guardianship reform, the prose says the reform increased the male-female labor-force participation gap by 7.4 percentage points. The table reports -7.38, with a note explicitly defining negative values as a closing gap. The prose reverses the direction of the displayed result.

GPT-4o mini misses that reversal. Instead, it flags a supposed mismatch between 1.1 in the prose and 1.10 in a table, although those numbers are identical. It also lists successful code execution and table reproduction as findings. None of the four returned items identifies the injected increase-versus-decrease contradiction.

Sign reversals account for 11 of its 45 scored prose-table misses; they are recurring, though numeric mismatches are more numerous. This case shows attention to presentation without a reliable check of meaning. It does not establish that every response follows the same pattern.

Based on the framework in *Verifying the Verifiers*. Scores follow CRED's published matcher, which can under-credit valid detections. Costs use dated price lists; comparisons are descriptive.
Live benchmark