FILING-EVIDENCE · RESULTS PER MODEL

OLMo-3-7B

AllenAI · 7.3B parameters · bf16

Krishna Nandula · Pramana Research 17 November 2026 Filing-Evidence v1.1

Status: part of an independent community research report, published by Pramana Research on 17 November 2026. It has **not been peer-reviewed** and has not been submitted to or accepted by any conference or journal. Corrections and feedback are welcome on GitHub.

In brief

17.1%
accuracy with checkable evidence, plain setup (rank 10 of 10)

17.3%
best setup: check + retry

70
wrong answers per 1,000 that passed every check (plain)

121/125
correct "not found" when the answer wasn't on the pages

Scorecard

Setup	Accuracy	Conf.-wrong	Abstention	Ign. scale	Tokens	Time
C1 · Plain	17.1% 14.8–19.8%	70 5.6–8.8%	121	21.4%	40	5.7 min
C2 · Check + retry	17.3% 14.9–19.9%	70 5.6–8.8%	123	21.5%	45	2.0 min
C3 · Thinking	n/a: this model has no reasoning mode					

Accuracy over the 875 answerable questions; confident-wrong per 1,000 answers; abstention out of 125; tokens are generated tokens per answer (mean). Intervals are 95% Wilson intervals. "Ign. scale" (ignoring scale) counts answers with the right digits and valid evidence

whatever scale was stated (a secondary metric). C2 tokens include the retry; C2 run time counts only the retries, because C2 reuses C1's first answers.

Where the answers go

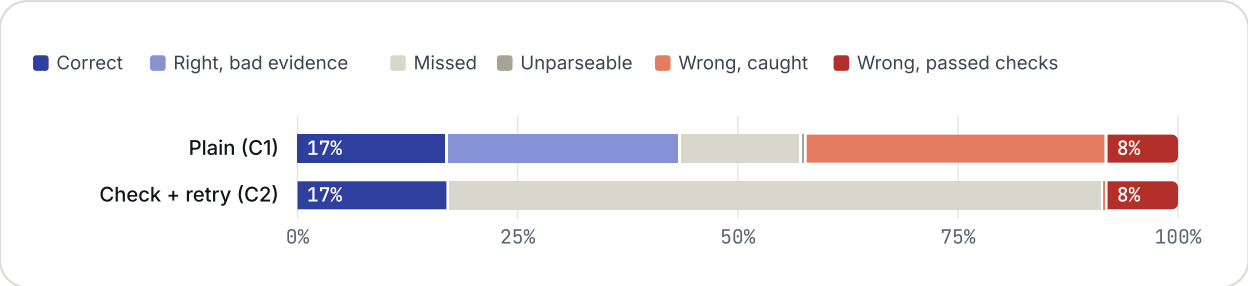

Figure 1. Outcomes of the 875 answerable questions in each setup.

In the plain setup, 150 answers were correct and 70 were wrong while passing every check: 37 scale or unit; 29 another period or column of the right line; 4 a different item. A further 231 had the right value with evidence the checker could not accept, and 299 wrong answers were rejected by the checker.

By kind of content

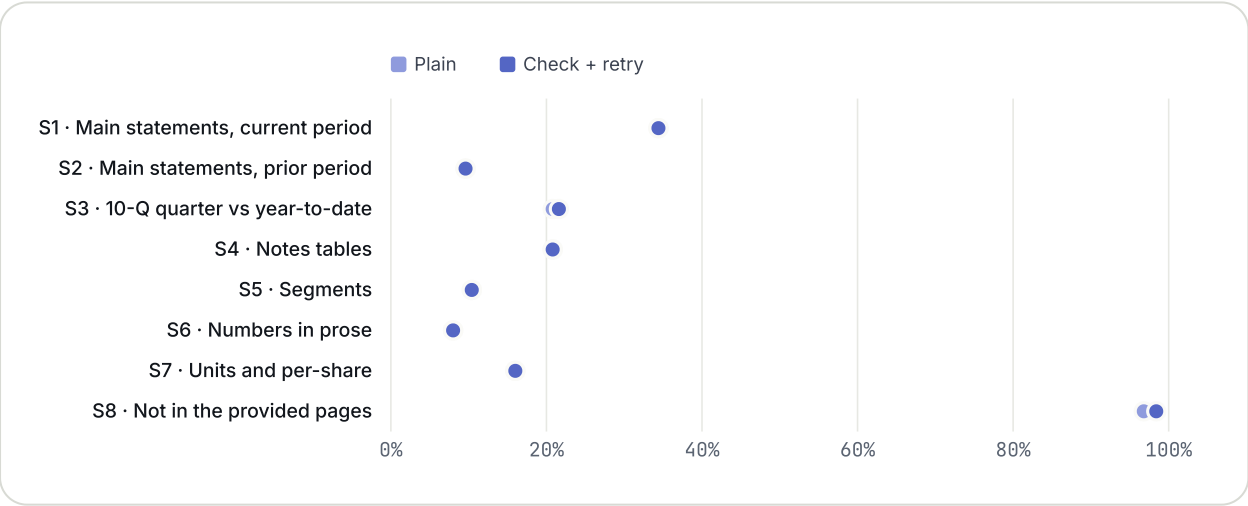

Figure 2. Accuracy per slice and setup (S8: correct "not found").

Plain setup: strongest on main statements, current period (34.4%) and notes tables (20.8%); weakest on numbers in prose (8.0%) and main statements, prior period (9.6%).

Retry and thinking

Comparison	Accuracy change	Holm p	Confident-wrong change	Holm p
Check + retry vs plain	+0.1 pts	1	+0	1

Paired exact McNemar tests on the same questions; Holm-corrected over every within-model comparison in the study. Blue: significant at α = 0.05.

Check + retry asked 537 answers again. Of those, 1 became correct, 0 became confident-wrong, 530 turned into "not found", and the rest still failed a check.

What stands out

- It usually finds the right number but rarely quotes the line exactly: it inserts dollar signs the page doesn't have ("26,873 $ 24,248"), repeats figures or joins lines. 231 answers had the right value with a quote the checker could not accept.
- Retry hardly helps (17.3%): asked again, it repeats the same kind of quote.

Real answers

Chosen by a fixed rule (evenly spaced over question ids) from its best setup (check + retry), not picked by hand.

Correct

American Express Company · 10-K · S1 Main statements, current period · FE-0015 CORRECT

“What was American Express Company's net increase in customer deposits for the fiscal year ended December 31, 2025?”

Answer	`13045` in millions, page 97
Its quote	`Net increase in customer deposits 13,045`
Answer key	`13,045` million · from `Net increase in customer deposits 13,045 10,305 18,915`

Thermo Fisher Scientific Inc. · 10-Q · S3 10-Q quarter vs year-to-date · FE-0352 CORRECT

"What was Thermo Fisher Scientific Inc.'s research and development expenses for the nine months ended September 26, 2026?"

Answer	`1048` in millions, page 3
Its quote	`Research and development expenses 348 346 1,048 1,041`
Answer key	`1,048` million · from `Research and development expenses 348 346 1,048 1,041`

Confident-wrong: passed every check, still wrong

Caterpillar Inc. · 10-K · S1 Main statements, current period · FE-0019 CONFIDENT-WRONG

"What was Caterpillar Inc.'s cash and cash equivalents as of December 31, 2025?"

Answer	`9980` in dollars, page 62
Its quote	`Cash and cash equivalents $ 9,980 $ 6,889`
Answer key	`9,980` million · from `Cash and cash equivalents $ 9,980 $ 6,889`
Error type	Scale or unit

Honeywell International Inc. · 10-Q · S2 Main statements, prior period · FE-0166 CONFIDENT-WRONG

"What was Honeywell International Inc.'s Less: Net income (loss) attributable to noncontrolling interest for the nine months ended September 30, 2025?"

Answer	`10` in millions, page 8
Its quote	`Less: Net income (loss) attributable to noncontrolling interest 12 34 (10) 51`
Answer key	`51` million · from `Less: Net income (loss) attributable to noncontrolling interest 12 34 (10) 51`
Error type	Wrong period or column

Medtronic plc · 10-Q · S4 Notes tables · FE-0440 CONFIDENT-WRONG

"What was Medtronic plc's income before income taxes for the three months ended July 25, 2025?"

Answer	`1302` none, page 34
Its quote	`Income before income taxes $ 1,302`
Answer key	`1,302` million · from `Income before income taxes $ 1,302`
Error type	Scale or unit

Wrong, caught by the checker

Johnson & Johnson · 10-Q · S3 10-Q quarter vs year-to-date · FE-0297 CAUGHT WRONG

"What was Johnson & Johnson's unrealized gain (loss) arising during period for the three months ended September 27, 2026?"

Answer	`58` in dollars, page 8
Its quote	`(Unrealized gain (loss) arising during period (58)`
Answer key	`58` million · from `Unrealized gain (loss) arising during period (58) 434 (274) 1,642`
Checker	failed the span check
Error type	Scale or unit

How it was run

Hugging Face repository
allenai/Olmo-3-7B-Instruct
(https://huggingface.co/allenai/Olmo-3-7B-Instruct)

Pinned revision
`6e5971d9eba4`

Licence
Apache 2.0

Released (repository created)
2025-11-19

Precision
bf16

Thinking switch
`none (C3 n/a)`

```
vllm serve allenai/Olmo-3-7B-Instruct --served-model-name olmo-3-7b --host
127.0.0.1 --port 8000 --max-model-len 16384 --gpu-memory-utilization 0.9 --
api-key <key> --seed 0 --revision 6e5971d9eba42665f5bd5a0fcf047f299ce1dccc
```

```
--chat-template research/benchmarks/filing-evidence/templates/olmo-3-
7b.jinja
```

```
uv run python hf_data.py pull
bash run_all.sh olmo-3-7b                                  # serve, then C1, C2
uv run python runner.py olmo-3-7b C1 --regrade          # re-grade the
stored replies
```

DISCLAIMER

This is an independent community research report, written in a personal capacity; it does not represent the views of any employer or other organisation. It is not affiliated with, sponsored or endorsed by any company or model developer named here; all names and trademarks belong to their owners. Results describe specific model versions under one fixed setup and are not a general judgement of any model. The data, results and code are provided as is, without warranty, and may contain errors; check any figure against the original filing. Nothing here is financial, investment or legal advice.

← All models

© 2026 Pramana Research · Text, data and figures: CC BY 4.0 · Code: Apache 2.0

Model outputs are not investment advice. Check every figure against its source.