FILING-EVIDENCE · RESULTS PER MODEL

Llama-3.1-8B

Meta · 8.0B 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 It has **not been peer-reviewed** and has not been submitted to or accepted by any conference or journal.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

58.9%
accuracy with checkable evidence, plain setup (rank 7 of 10)

62.7%
best setup: check + retry

102
wrong answers per 1,000 that passed every check (plain)

110/125
correct "not found" when the answer wasn't on the pages

Scorecard

Setup	Accuracy	Conf.-wrong	Abstention	Ign. scale	Tokens	Time
C1 · Plain	58.9% 55.6–62.1%	102 8.5–12.2%	110	65.9%	49	4.4 min
C2 · Check + retry	62.7% 59.5–65.9%	124 10.5–14.6%	110	70.7%	65	1.3 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

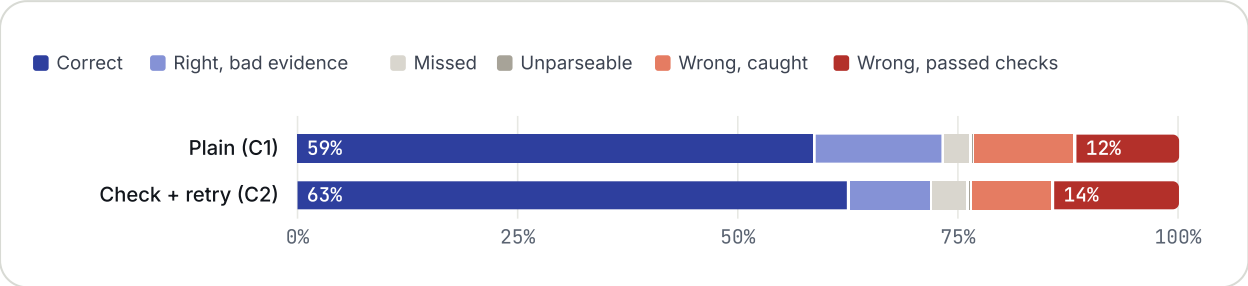

Figure 1. Outcomes of the 875 answerable questions in each setup.

In the plain setup, 515 answers were correct and 102 were wrong while passing every check: 62 scale or unit; 23 another period or column of the right line; 17 a different item. A further 128 had the right value with evidence the checker could not accept, and 101 wrong answers were rejected by the checker.

By kind of content

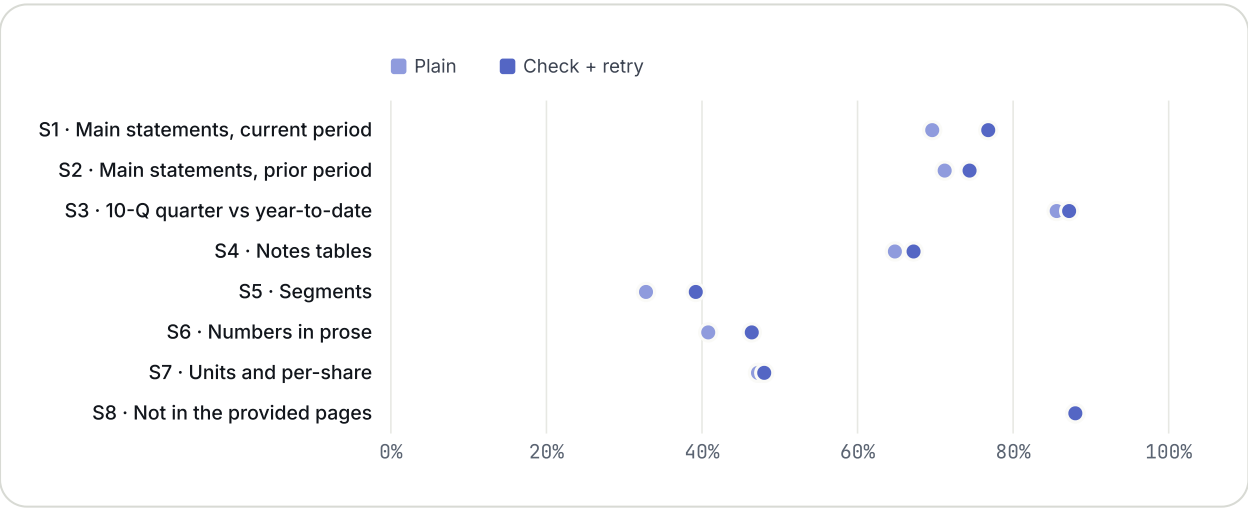

Figure 2. Accuracy per slice and setup (S8: correct "not found").

Plain setup: strongest on 10-Q quarter vs year-to-date (85.6%) and main statements, prior period (71.2%); weakest on segments (32.8%) and numbers in prose (40.8%).

Retry and thinking

Comparison	Accuracy change	Holm p	Confident-wrong change	Holm p
Check + retry vs plain	+3.9 pts	2.56e-09	+22	8.11e-06

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 233 answers again. Of those, 34 became correct, 22 became confident-wrong, 9 turned into "not found", and the rest still failed a check.

What stands out

- The oldest model in the study (released 2024-07), statistically level with Gemma-3-12B and Phi-4.
- Retry raised its confident-wrong answers the most of any model, from 102 to 124.

Real answers

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

Correct

Abbott Laboratories · 10-Q · S1 Main statements, current period · FE-0005 CORRECT

“What was Abbott Laboratories's Total Current Liabilities as of September 30, 2026?”

Answer	`16152` in millions, page 4
Its quote	`Current Liabilities: Trade accounts payable $ 4,656 $ 4,240 Salaries, wages, and commissions 1,896 1,745 Other accrued liabilities 6,583 5,812 Dividends payable 1,094 1,097 Income taxes payable 215 569 Current portion of long-term debt 1,708 3,033 Total Current Liabilities 16,152 16,496`
Answer key	`16,152` million · from `Total Current Liabilities 16,152 16,496`

Simon Property Group, Inc. · 10-Q · S3 10-Q quarter vs year-to-date · FE-0349 CORRECT

"What was Simon Property Group, Inc.'s consolidated net income for the three months ended September 30, 2026?"

Answer	`565294` in thousands, page 5
Its quote	`CONSOLIDATED NET INCOME 565,294 702,696 1,707,959 1,824,237`
Answer key	`565,294` thousand · from `CONSOLIDATED NET INCOME 565,294 702,696 1,707,959 1,824,237`

Confident-wrong: passed every check, still wrong

Abbott Laboratories · 10-Q · S1 Main statements, current period · FE-0006
CONFIDENT-WRONG

"What was Abbott Laboratories's net property and equipment as of September 30, 2026?"

Answer	`12880` none, page 4
Its quote	`Net property and equipment 12,880 11,816`
Answer key	`12,880` million · from `Net property and equipment 12,880 11,816`
Error type	Scale or unit

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

Wells Fargo & Company · 10-Q · S6 Numbers in prose · FE-0748 CONFIDENT-WRONG

“What was Wells Fargo & Company's net derivative gains (losses) from economic hedges related to mortgage loans held for sale and derivative loan commitments for the three months ended September 30, 2026?”

Answer	`-390` in millions, page 83
Its quote	`Net derivative gains (losses) from economic hedges (4) (390) 2 (471) 128`
Answer key	`87` million · from `(6)Includes net gains (losses) of $87 million and $116 million in the third quarter and first nine months of 2026, respectively, and $(6) million and $(20) million in the third quarter and first nine months of 2025, respectively, related to derivatives used as economic hedges of mortgage loans held for sale and derivative loan commitments.`
Error type	Wrong item

Right value, unusable evidence

Progressive Corporation · 10-Q · S1 Main statements, current period · FE-0088
RIGHT VALUE, UNUSABLE EVIDENCE

“What was Progressive Corporation's net unrealized gains (losses) on fixed-maturity securities as of September 30, 2026?”

Answer	`-2730` in millions, page 43
Its quote	`The decline from September 30, 2025 and December 31, 2025 was due to valuation decreases across fixed-maturity sectors as interest rates rose during 2026. Our U.S. government and corporate and other debt securities had the most significant valuation decrease from prior year and prior year end.`
Answer key	`2,730` million · from `Net unrealized gains (losses) on fixed-maturity securities (2,730) 203 117`
Checker	failed the value check
Error type	Wrong line quoted

Wrong, caught by the checker

United Parcel Service, Inc. · 10-Q · S1 Main statements, current period · FE-0114

CAUGHT WRONG

"What was United Parcel Service, Inc.'s Pension and Postretirement Benefit Obligations as of September 30, 2026?"

Answer	`5747` none, page 3
Its quote	`Pension and Postretirement Benefit Obligations 5,751 6,567`
Answer key	`5,751` million · from `Pension and Postretirement Benefit Obligations 5,751 6,567`
Checker	failed the value check
Error type	Wrong item

How it was run

Hugging Face repository
meta-llama/Llama-3.1-8B-Instruct
(https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct)

Pinned revision
`0e9e39f249a1`
reconstructed from the local cache: the server was started without the pin (PROTOCOL §12)

Licence
Llama 3.1 Community

Released (repository created)
2024-07-18

Precision
bf16

Thinking switch
`none (C3 n/a)`

```
vllm serve meta-llama/Llama-3.1-8B-Instruct --served-model-name llama-3.1-8b --host 127.0.0.1 --port 8000 --max-model-len 16384 --gpu-memory-utilization 0.9 --api-key <key> --seed 0
```

```
uv run python hf_data.py pull
bash run_all.sh llama-3.1-8b                                  # serve, then C1, C2
uv run python runner.py llama-3.1-8b 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.