FILING-EVIDENCE · RESULTS PER MODEL

Gemma-3-12B

Google · 12.2B parameters · fp8

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 It has **not been peer-reviewed** and has not been submitted to or accepted by any conference or journal.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

61.4%
accuracy with checkable evidence, plain setup (rank 6 of 10)

64.3%
best setup: check + retry

117
wrong answers per 1,000 that passed every check (plain)

97/125
correct "not found" when the answer wasn't on the pages

Scorecard

Setup	Accuracy	Conf.-wrong	Abstention	Ign. scale	Tokens	Time
C1 · Plain	61.4% 58.1–64.5%	117 9.9–13.8%	97	68.7%	61	6.5 min
C2 · Check + retry	64.3% 61.1–67.4%	134 11.4–15.7%	98	72.0%	74	1.5 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

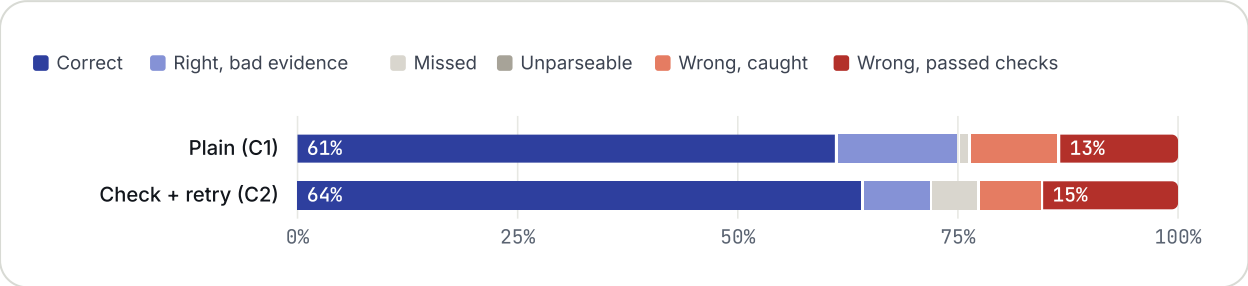

Figure 1. Outcomes of the 875 answerable questions in each setup.

In the plain setup, 537 answers were correct and 117 were wrong while passing every check: 64 scale or unit; 29 another period or column of the right line; 24 a different item. A further 121 had the right value with evidence the checker could not accept, and 88 wrong answers were rejected by the checker.

By kind of content

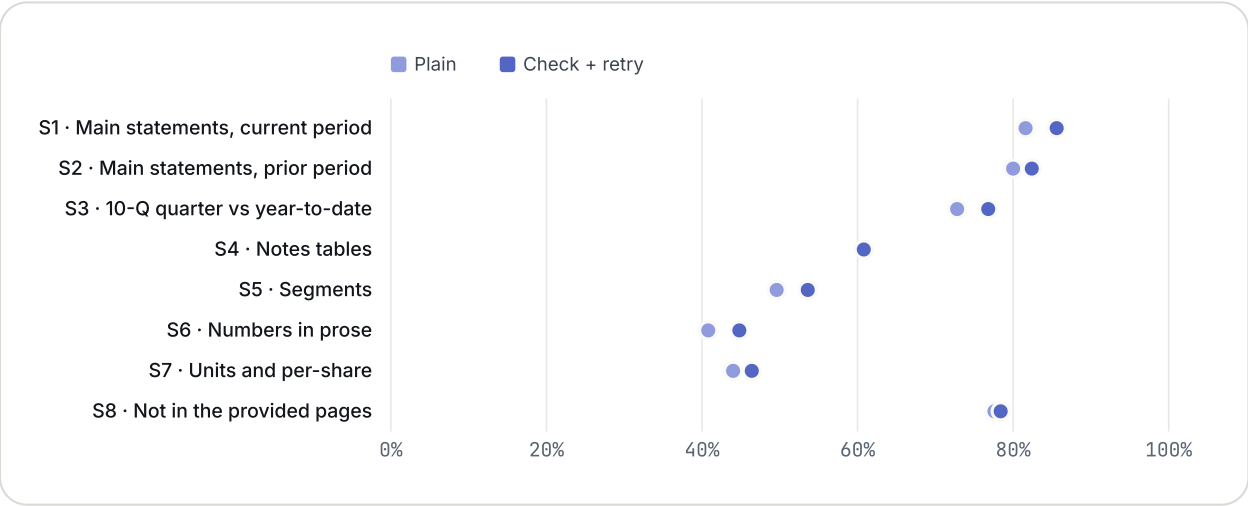

Figure 2. Accuracy per slice and setup (S8: correct "not found").

Plain setup: strongest on main statements, current period (81.6%) and main statements, prior period (80.0%); weakest on numbers in prose (40.8%) and units and per-share (44.0%).

Retry and thinking

Comparison	Accuracy change	Holm p	Confident-wrong change	Holm p
Check + retry vs plain	+3.0 pts	4.77e-07	+17	0.000244

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 214 answers again. Of those, 26 became correct, 17 became confident-wrong, 36 turned into "not found", and the rest still failed a check.

What stands out

- The weakest of the ten at saying "not found" (97 of 125): it most often answers a quarterly figure asked of an annual report, where the correct answer is "not found" (17 of 31 correct).
- Run with FP8 weights; retry raised its confident-wrong answers from 117 to 134.

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	`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	`Income from unconsolidated entities 152,992 143,916 250,871 297,150 Unrealized (losses) gains in fair value of publicly traded equity instruments and derivative instrument, net (129,565) 2,243 (160,602) (84,977) Gain on acquisition of controlling interest, sale or disposal of, or recovery on, assets and interests in unconsolidated entities and impairment, net 10,701 10,398 66,294 794 CONSOLIDATED NET INCOME 565,294`
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

U.S. Bancorp · 10-Q · S3 10-Q quarter vs year-to-date · FE-0364 CONFIDENT-WRONG

"What was U.S. Bancorp's net interest income after provision for credit losses for the three months ended September 30, 2026?"

Answer	`11324` in millions, page 31
Its quote	`Net interest income after provision for credit losses 3,814 3,651 11,324 10,756`
Answer key	`3,814` million · from `Net interest income after provision for credit losses 3,814 3,651 11,324 10,756`
Error type	Wrong period or column

Pfizer Inc. · 10-Q · S6 Numbers in prose · FE-0707 CONFIDENT-WRONG

“What was Pfizer Inc.'s long-term equity securities held in trust as of December 31, 2025?”

Answer	`642` in millions, page 16
Its quote	`Equity securities with readily determinable fair values(a) $ 484 $ 642`
Answer key	`146` million · from `(b)Long-term equity securities of $122 million as of September 27, 2026 and $146 million as of December 31, 2025 were held in restricted trusts for U.S. non-qualified employee benefit plans.`
Error type	Wrong item

Right value, unusable evidence

Coca-Cola Company · 10-K · S2 Main statements, prior period · FE-0174
RIGHT VALUE, UNUSABLE EVIDENCE

“What was Coca-Cola Company's deferred income taxes for the fiscal year ended December 31, 2024?”

Answer	`-11` in millions, page 65
Its quote	`Deferred income taxes (11)`
Answer key	`11` million · from `Deferred income taxes 517 (11) (2)`
Checker	failed the span check
Error type	Quote not verbatim

Wrong, caught by the checker

U.S. Bancorp · 10-Q · S2 Main statements, prior period · FE-0240 CAUGHT WRONG

"What was U.S. Bancorp's net (income) loss attributable to noncontrolling interests for the three months ended September 30, 2025?"

Answer	`7` none, page 32
Its quote	`Comprehensive income (loss) attributable to noncontrolling interests (6) (7)`
Answer key	`7` million · from `Net (income) loss attributable to noncontrolling interests (6) (7) (17) (20)`
Checker	failed the span check
Error type	Scale or unit

How it was run

Hugging Face repository
google/gemma-3-12b-it
(https://huggingface.co/google/gemma-3-12b-it)

Pinned revision
`96b6f1eccf38`

Licence
Gemma Terms of Use

Released (repository created)
2025-03-01

Precision
fp8

Thinking switch
`none (C3 n/a)`

```
vllm serve google/gemma-3-12b-it --served-model-name gemma-3-12b --host
127.0.0.1 --port 8000 --max-model-len 16384 --gpu-memory-utilization 0.9 --
api-key <key> --seed 0 --revision 96b6f1eccf38110c56df3a15bffe176da04bfd80
--quantization fp8
```

```
uv run python hf_data.py pull
bash run_all.sh gemma-3-12b                                  # serve, then C1,
C2
uv run python runner.py gemma-3-12b 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.