FILING-EVIDENCE · RESULTS PER MODEL

Ministral-3-8B

Mistral · 8.9B 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 It has **not been peer-reviewed** and has not been submitted to or accepted by any conference or journal.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

66.7%
accuracy with checkable evidence, plain setup (rank 5 of 10)

69.6%
best setup: check + retry

77
wrong answers per 1,000 that passed every check (plain)

124/125
correct "not found" when the answer wasn't on the pages

Scorecard

Setup	Accuracy	Conf.-wrong	Abstention	Ign. scale	Tokens	Time
C1 · Plain	66.7% 63.6–69.8%	77 6.2–9.5%	124	72.8%	64	4.0 min
C2 · Check + retry	69.6% 66.5–72.6%	86 7.0–10.5%	124	75.7%	77	0.8 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

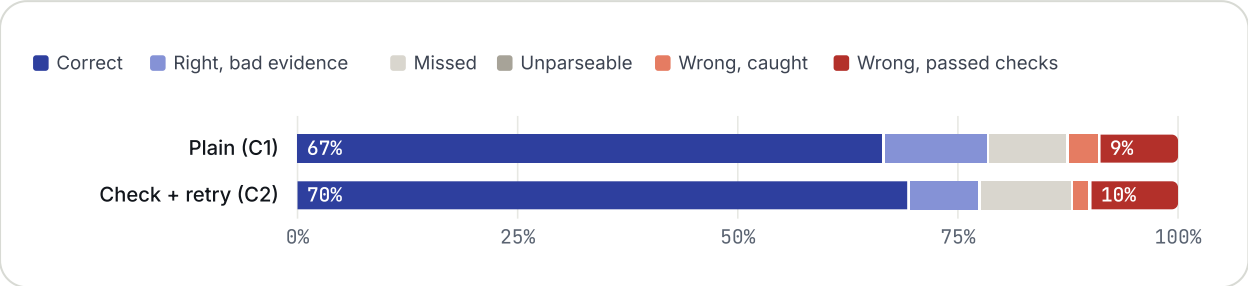

Figure 1. Outcomes of the 875 answerable questions in each setup.

In the plain setup, 584 answers were correct and 77 were wrong while passing every check: 53 scale or unit; 16 another period or column of the right line; 8 a different item. A further 104 had the right value with evidence the checker could not accept, and 31 wrong answers were rejected by the checker.

By kind of content

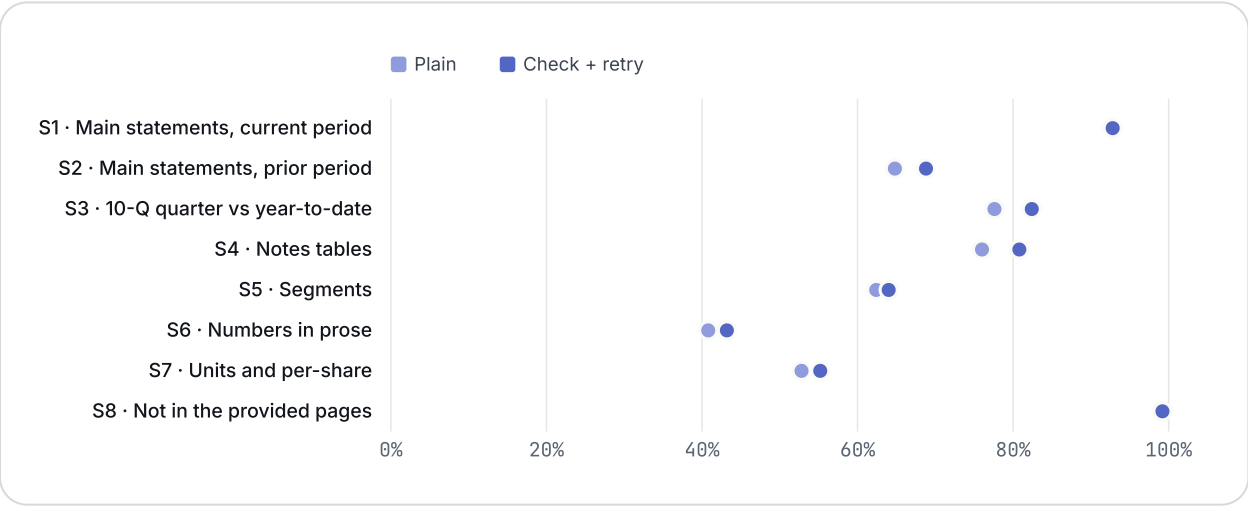

Figure 2. Accuracy per slice and setup (S8: correct "not found").

Plain setup: strongest on main statements, current period (92.8%) and 10-Q quarter vs year-to-date (77.6%); weakest on numbers in prose (40.8%) and units and per-share (52.8%).

Retry and thinking

Comparison	Accuracy change	Holm p	Confident-wrong change	Holm p
Check + retry vs plain	+2.9 pts	7.75e-07	+9	0.0422

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 135 answers again. Of those, 25 became correct, 9 became confident-wrong, 13 turned into "not found", and the rest still failed a check.

What stands out

- Among the best at saying "not found" (124 of 125).
- No reasoning mode. Retry adds 2.9 points of accuracy but also 9 confident-wrong answers.

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`

Apple Inc. · 10-K · S4 Notes tables · FE-0377 CORRECT

"What was Apple Inc.'s capitalized research and development as of September 27, 2025?"

Answer	`15041` in millions, page 47
Its quote	`Capitalized research and development $ 13,656 $ 15,041`
Answer key	`15,041` million · from `Capitalized research and development $ 13,656 $ 15,041`

Confident-wrong: passed every check, still wrong

Pfizer Inc. · 10-K · S1 Main statements, current period · FE-0085 CONFIDENT-WRONG

"What was Pfizer Inc.'s net increase/(decrease) in cash and cash equivalents and restricted cash and cash equivalents for the fiscal year ended December 31, 2025?"

Answer	`91` none, page 61
Its quote	`Net increase/(decrease) in cash and cash equivalents and restricted cash and cash equivalents 91 (1,810) 2,448`
Answer key	`91` million · from `Net increase/(decrease) in cash and cash equivalents and restricted cash and cash equivalents 91 (1,810) 2,448`
Error type	Scale or unit

Walt Disney Company · 10-Q · S5 Segments · FE-0528 CONFIDENT-WRONG

"What was Walt Disney Company's goodwill for its Experiences Segment as of September 27, 2025?"

Answer	`51258` in millions, page 15
Its quote	`Balance at September 27, 2025 $ 51,258 $ 16,486 $ 5,550 $ 73,294`
Answer key	`5,550` million · from `Balance at September 27, 2025 $ 51,258 $ 16,486 $ 5,550 $ 73,294`
Error type	Wrong period or column

McDonald's Corporation · 10-Q · S7 Units and per-share · FE-0808 CONFIDENT-WRONG

"What was McDonald's Corporation's weighted-average shares outstanding-basic (number) for the nine months ended September 30, 2025?"

Answer	`712.9` millions, page 3
Its quote	`Weighted-average shares outstanding-basic 706.6 712.9 708.8 714.1`
Answer key	`714.1` · from `Weighted-average shares outstanding-basic 706.6 712.9 708.8 714.1`
Error type	Wrong period or column

Right value, unusable evidence

Apple Inc. · 10-K · S2 Main statements, prior period · FE-0128
RIGHT VALUE, UNUSABLE EVIDENCE

"What was Apple Inc.'s proceeds from sales of marketable securities for the fiscal year ended September 28, 2024?"

Answer	`11135` in millions, page 37
Its quote	`Proceeds from sales of marketable securities 11,135`
Answer key	`11,135` million · from `Proceeds from sales of marketable securities 14,280 12,890 11,135`
Checker	failed the span check
Error type	Quote not verbatim

Wrong, caught by the checker

Medtronic plc · 10-Q · S5 Segments · FE-0570 CAUGHT WRONG

"What was Medtronic plc's goodwill for its Medical Surgical segment as of July 31, 2026?"

Answer	`19953` in millions, page 26
Its quote	`Medical Surgical $ 19,953 $ 40,332 $ 2,256 $ 42,587`
Answer key	`19,675` million · from `July 31, 2026 $ 8,583 $ 12,673 $ 19,675 $ 40,932 $ 2,255 $ 43,187`
Checker	failed the span check
Error type	Wrong item

How it was run

Hugging Face repository
mistralai/Ministral-3-8B-Instruct-2512
(https://huggingface.co/mistralai/Ministral-3-8B-Instruct-2512)

Pinned revision
`5b26027e7b19`

Licence
Apache 2.0

Released (repository created)
2025-10-31

Precision
bf16

Thinking switch
`none (C3 n/a)`

```
vllm serve mistralai/Ministral-3-8B-Instruct-2512 --served-model-name
ministral-3-8b --host 127.0.0.1 --port 8000 --max-model-len 16384 --gpu-
memory-utilization 0.9 --api-key <key> --seed 0 --revision
5b26027e7b19eeb4b7352e1fed3926375dd2cb4d --tokenizer-mode mistral --config-
format mistral --load-format mistral
```

```
uv run python hf_data.py pull
bash run_all.sh ministral-3-8b                                  # serve, then
C1, C2
uv run python runner.py ministral-3-8b 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.