FILING-EVIDENCE · RESULTS PER MODEL

Granite-4.2-8B

IBM · 8.8B parameters · bf16 · has a thinking mode

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 It has **not been peer-reviewed** and has not been submitted to or accepted by any conference or journal.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

73.3%
accuracy with checkable evidence, plain setup (rank 3 of 10)

81.3%
best setup: thinking

64
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	73.3% 70.2–76.1%	64 5.0–8.1%	121	75.1%	49	5.5 min
C2 · Check + retry	76.2% 73.3–78.9%	72 5.8–9.0%	121	78.3%	58	0.9 min
C3 · Thinking	81.3% 78.5–83.7%	38 2.8–5.2%	123	83.7%	2,538	182.1 min

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

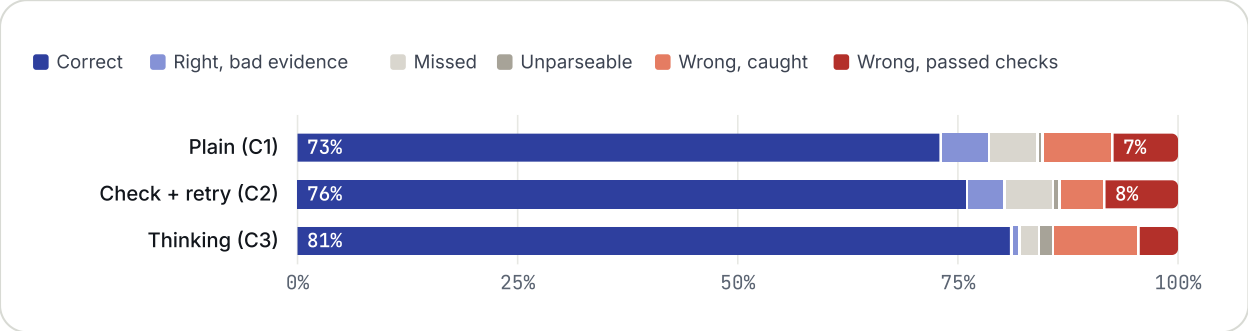

Figure 1. Outcomes of the 875 answerable questions in each setup.

In the plain setup, 641 answers were correct and 64 were wrong while passing every check: 16 scale or unit; 35 another period or column of the right line; 13 a different item. A further 48 had the right value with evidence the checker could not accept, and 69 wrong answers were rejected by the checker.

By kind of content

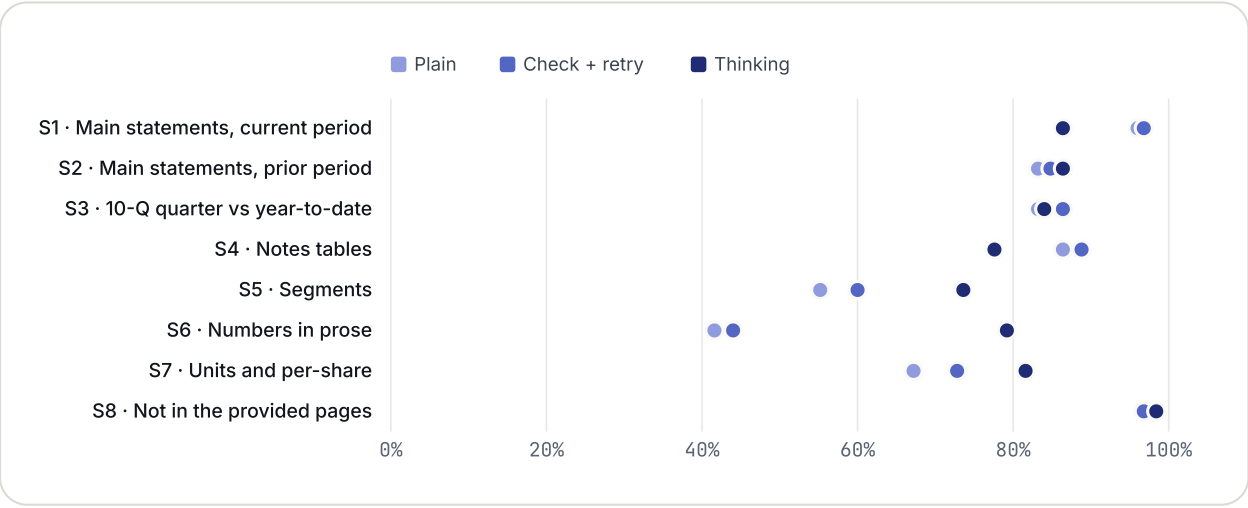

Figure 2. Accuracy per slice and setup (S8: correct "not found").

Plain setup: strongest on main statements, current period (96.0%) and notes tables (86.4%); weakest on numbers in prose (41.6%) and segments (55.2%).

Retry and thinking

Comparison	Accuracy change	Holm p	Confident-wrong change	Holm p
Check + retry vs plain	+3.0 pts	4.77e-07	+8	0.0579
Thinking vs plain	+8.0 pts	0.00011	-26	0.0422
Thinking vs check + retry	+5.0 pts	0.0201	-34	0.00209

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 122 answers again. Of those, 26 became correct, 8 became confident-wrong, 0 turned into "not found", and the rest still failed a check.

What stands out

- Second-fewest confident-wrong answers among the ten in the plain setup (64), and 38 with thinking.
- Thinking raises accuracy to 81.3%, statistically level with the best thinking-mode results, at about 52 times the tokens.
- It thinks by default; the plain setup switches thinking off through its chat template.

Real answers

Chosen by a fixed rule (evenly spaced over question ids) from its best setup (thinking), 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`

Coca-Cola Company · 10-K · S4 Notes tables · FE-0426 CORRECT

"What was Coca-Cola Company's total income taxes paid, net of refunds for the fiscal year ended December 31, 2025?"

Answer	`2873` in millions, page 103
Its quote	`Total income taxes paid, net of refunds $ 2,873`
Answer key	`2,873` million · from `Total income taxes paid, net of refunds $ 2,873`

Confident-wrong: passed every check, still wrong

Bank of America Corporation · 10-Q · S2 Main statements, prior period · FE-0140
CONFIDENT-WRONG

"What was Bank of America Corporation's net change in derivatives for the nine months ended September 30, 2025?"

Answer	`636` in millions, page 48
Its quote	`Net change in derivatives (1,496) 636 (2,874) 3,145`
Answer key	`3,145` million · from `Net change in derivatives (1,496) 636 (2,874) 3,145`
Error type	Wrong period or column

Duke Energy Corporation · 10-K · S5 Segments · FE-0531 CONFIDENT-WRONG

"What was Duke Energy Corporation's investments in equity method unconsolidated affiliates for its other segment as of December 31, 2024?"

Answer	`36` in millions, page 182
Its quote	`Other 153 36 139 50 66`
Answer key	`139` million · from `Other 153 36 139 50 66`
Error type	Wrong period or column

Honeywell International Inc. · 10-Q · S7 Units and per-share · FE-0791 CONFIDENT-WRONG

"What was Honeywell International Inc.'s (Loss) earnings per share of common stock from discontinued operations—assuming dilution for the three months ended September 30, 2025?"

Answer	`(0.32)` none, page 33
Its quote	`(Loss) earnings per share of common stock from discontinued operations—assuming dilution (0.32) 2.72 3.27 10.05`
Answer key	`2.72` · from `(Loss) earnings per share of common stock from discontinued operations—assuming dilution (0.32) 2.72 3.27 10.05`
Error type	Wrong period or column

Right value, unusable evidence

Prologis, Inc. · 10-Q · S3 10-Q quarter vs year-to-date · FE-0333
RIGHT VALUE, UNUSABLE EVIDENCE

"What was Prologis, Inc.'s earnings before income taxes for the three months ended September 30, 2026?"

Answer	`977861` in thousands, page 4
Its quote	`Earnings before income taxes $ 977,861 $ 875,795 $ 3,302,774 $ 2,204,407`
Answer key	`977,861` thousand · from `Earnings before income taxes 977,861 875,795 3,302,774 2,204,407`
Checker	failed the span check
Error type	Wrong page cited

Wrong, caught by the checker

Bank of America Corporation · 10-K · S1 Main statements, current period · FE-0018

CAUGHT WRONG

“What was Bank of America Corporation's net cash used in investing activities for the fiscal year ended December 31, 2025?”

Answer	`2511` in millions, page 12
Its quote	`Cost of sales 2,605 2,511`
Answer key	`145,157` million · from `Net cash used in investing activities (145,157) (90,693) (35,387)`
Checker	failed the page check
Error type	Wrong item

How it was run

Hugging Face repository ibm-granite/granite-4.2-8b (https://huggingface.co/ibm-granite/granite-4.2-8b)	Pinned revision `a7042144125c`
Licence Apache 2.0	Released (repository created) 2026-08-07
Precision bf16	Thinking switch `{'chat_template_kwargs': {'enable_thinking': True}}`

```
vllm serve ibm-granite/granite-4.2-8b --served-model-name granite-4.2-8b --
host 127.0.0.1 --port 8000 --max-model-len 16384 --gpu-memory-utilization
0.9 --api-key <key> --seed 0 --revision
a7042144125ce5ae64f89610e8289647fc000b9a
```

```
uv run python hf_data.py pull
bash run_all.sh granite-4.2-8b                                  # serve, then
C1, C2, C3
uv run python runner.py granite-4.2-8b 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.