FILING-EVIDENCE · RESULTS PER MODEL

gpt-oss-20b

OpenAI · 20.9B parameters · mxfp4 · 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 It has **not been peer-reviewed** and has not been submitted to or accepted by any conference or journal.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

74.4%
accuracy with checkable evidence, plain setup (rank 2 of 10)

80.7%
best setup: thinking

112
wrong answers per 1,000 that passed every check (plain)

112/125
correct "not found" when the answer wasn't on the pages

Scorecard

Setup	Accuracy	Conf.-wrong	Abstention	Ign. scale	Tokens	Time
C1 · Plain	74.4% 71.4–77.2%	112 9.4–13.3%	112	80.9%	109	3.3 min
C2 · Check + retry	77.4% 74.5–80.0%	118 9.9–13.9%	112	84.2%	120	0.4 min
C3 · Thinking	80.7% 77.9–83.2%	56 4.3–7.2%	106	84.5%	1,904	32.1 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

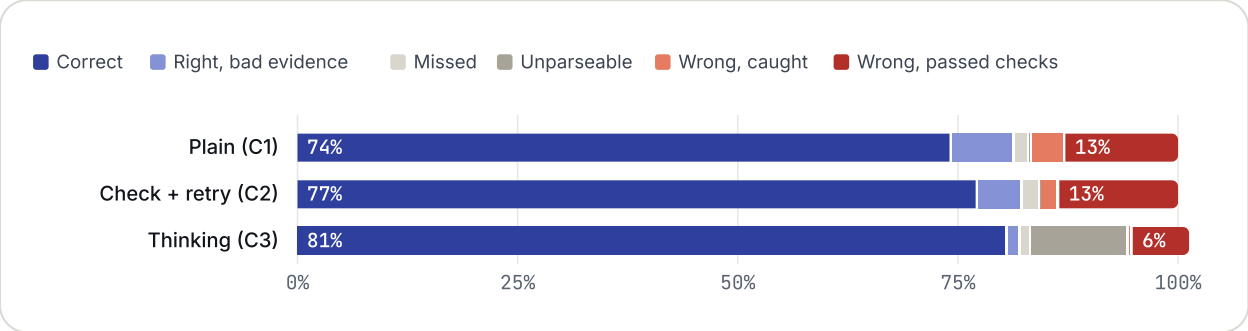

Figure 1. Outcomes of the 875 answerable questions in each setup.

In the plain setup, 651 answers were correct and 112 were wrong while passing every check: 57 scale or unit; 45 another period or column of the right line; 10 a different item. A further 62 had the right value with evidence the checker could not accept, and 33 wrong answers were rejected by the checker.

By kind of content

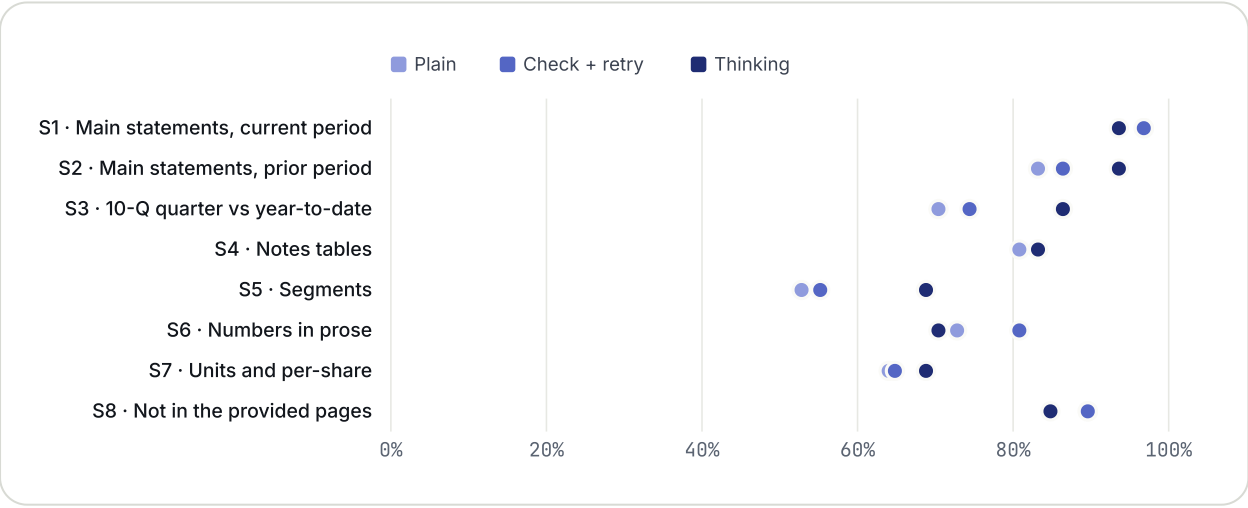

Figure 2. Accuracy per slice and setup (S8: correct "not found").

Plain setup: strongest on main statements, current period (96.8%) and main statements, prior period (83.2%); weakest on segments (52.8%) and units and per-share (64.0%).

Retry and thinking

Comparison	Accuracy change	Holm p	Confident-wrong change	Holm p
Check + retry vs plain	+3.0 pts	**4.77e-07**	+6	0.156
Thinking vs plain	+6.3 pts	**0.000462**	-56	**1.32e-07**
Thinking vs check + retry	+3.3 pts	0.105	-62	**2.09e-09**

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 99 answers again. Of those, 26 became correct, 6 became confident-wrong, 3 turned into "not found", and the rest still failed a check.

What stands out

- It has no non-reasoning mode, so its plain and check + retry setups use the lowest reasoning effort ("low") and its thinking setup the highest ("high"). Its plain-setup numbers are not directly comparable with models answering without reasoning.
- High effort halves its confident-wrong answers (112 → 56) but hit the 8,192-token limit on 97 questions.
- The largest model in the study (20.9B, mixture of experts), run in its native MXFP4 weights.

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-0004 CORRECT

“What was Abbott Laboratories's cumulative gains (losses) on derivative instruments designated as cash flow hedges as of September 30, 2026?”

Answer	`61` in millions, page 3
Its quote	`Cumulative gains (losses) on derivative instruments designated as cash flow hedges 61 (69)`
Answer key	`61` million · from `Cumulative gains (losses) on derivative instruments designated as cash flow hedges 61 (69)`

Allstate Corporation · 10-K · S4 Notes tables · FE-0385 CORRECT

"What was Allstate Corporation's net income (loss) for the fiscal year ended December 31, 2025?"

Answer	`10266` in millions, page 102
Its quote	`Net income (loss) $ 10,266 $ 4,599 $ (213)`
Answer key	`10,266` million · from `Net income (loss) $ 10,266 $ 4,599 $ (213)`

Confident-wrong: passed every check, still wrong

Adobe Inc. · 10-Q · S3 10-Q quarter vs year-to-date · FE-0260 CONFIDENT-WRONG

"What was Adobe Inc.'s unrealized gains / losses on derivative instruments for the nine months ended August 28, 2026?"

Answer	`39` in millions, page 4
Its quote	`Unrealized gains / losses on derivative instruments 39 (44) (3) (231)`
Answer key	`3` million · from `Unrealized gains / losses on derivative instruments 39 (44) (3) (231)`
Error type	Wrong period or column

UnitedHealth Group Incorporated · 10-Q · S5 Segments · FE-0614 CONFIDENT-WRONG

"What was UnitedHealth Group Incorporated's Depreciation and Amortization for its optum segment for the three months ended September 30, 2026?"

Answer	`195` in millions, page 16
Its quote	`Depreciation and Amortization 195 253 439 186 — 878 — 1,073`
Answer key	`878` million · from `Depreciation and Amortization 195 253 439 186 — 878 — 1,073`
Error type	Wrong period or column

NVIDIA Corporation · 10-K · S7 Units and per-share · FE-0822 CONFIDENT-WRONG

"What was NVIDIA Corporation's basic weighted average shares (number) for the fiscal year ended January 26, 2025?"

Answer	`24555` none, page 62
Its quote	`Basic weighted average shares 24,359 24,555 24,690`
Answer key	`24,555` million · from `Basic weighted average shares 24,359 24,555 24,690`
Error type	Scale or unit

Right value, unusable evidence

McdDonald's Corporation · 10-Q · S3 10-Q quarter vs year-to-date · FE-0313
RIGHT VALUE, UNUSABLE EVIDENCE

"What was McDonald's Corporation's cash and equivalents increase (decrease) for the nine months ended September 30, 2026?"

Answer	`419` in millions, page 5
Its quote	`Cash and equivalents increase (decrease) 419 1,328`
Answer key	`419` million · from `Cash and equivalents increase (decrease) 371 537 419 1,328`
Checker	failed the span check
Error type	Quote not verbatim

Wrong, caught by the checker

Lockheed Martin Corporation · 10-K · S5 Segments · FE-0561 CAUGHT WRONG

"What was Lockheed Martin Corporation's goodwill for its Space Systems segment as of December 31, 2024?"

Answer	`6712` in millions, page 73
Its quote	`Balance at December 31, 2024 196 2,085 6,712 2,074 11,067`
Answer key	`2,074` million · from `Balance at December 31, 2024 196 2,085 6,712 2,074 11,067`
Checker	failed the span check
Error type	Wrong period or column

How it was run

Hugging Face repository openai/gpt-oss-20b (https://huggingface.co/openai/gpt-oss-20b)	Pinned revision 6cee5e81ee83 reconstructed from the local cache: the server was started without the pin (PROTOCOL §12)
Licence Apache 2.0	Released (repository created) 2025-08-04
Precision mxfp4	Thinking switch {'reasoning_effort': 'high'}

```
vllm serve openai/gpt-oss-20b --served-model-name gpt-oss-20b --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 gpt-oss-20b                                  # serve, then C1,
C2, C3
uv run python runner.py gpt-oss-20b 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.