Qwen3-8B

Alibaba · 8.2B 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 It has **not been peer-reviewed** and has not been submitted to or accepted by any conference or journal.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

66.1%
accuracy with checkable evidence, plain setup (rank — of 10; not ranked)

74.1%
best setup: thinking

94
wrong answers per 1,000 that passed every check (plain)

88/125
correct "not found" when the answer wasn't on the pages

Scorecard

Setup	Accuracy	Conf.-wrong	Abstention	Ign. scale	Tokens	Time
C1 · Plain	66.1% 62.9–69.1%	94 7.7–11.4%	88	70.7%	59	5.5 min
C2 · Check + retry	67.3% 64.1–70.3%	101 8.4–12.1%	90	72.2%	72	1.2 min
C3 · Thinking	74.1% 71.1–76.9%	37 2.7–5.1%	104	75.5%	969	48.4 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

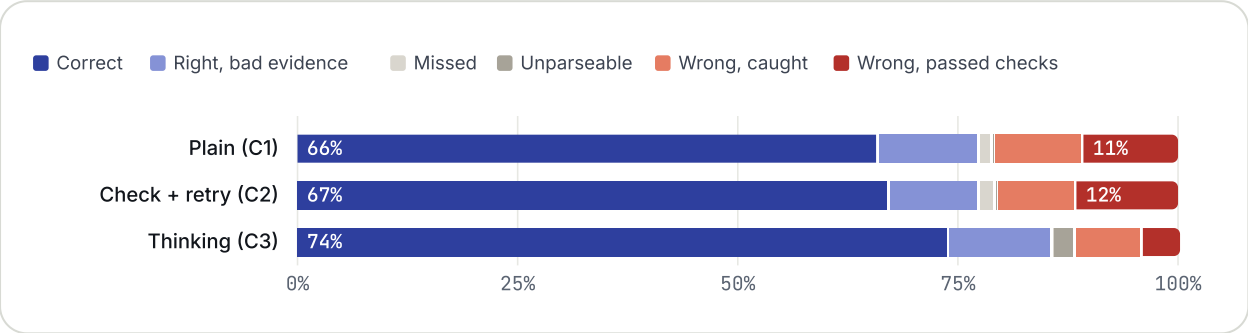

Figure 1. Outcomes of the 875 answerable questions in each setup.

In the plain setup, 578 answers were correct and 94 were wrong while passing every check: 41 scale or unit; 37 another period or column of the right line; 16 a different item. A further 100 had the right value with evidence the checker could not accept, and 88 wrong answers were rejected by the checker.

By kind of content

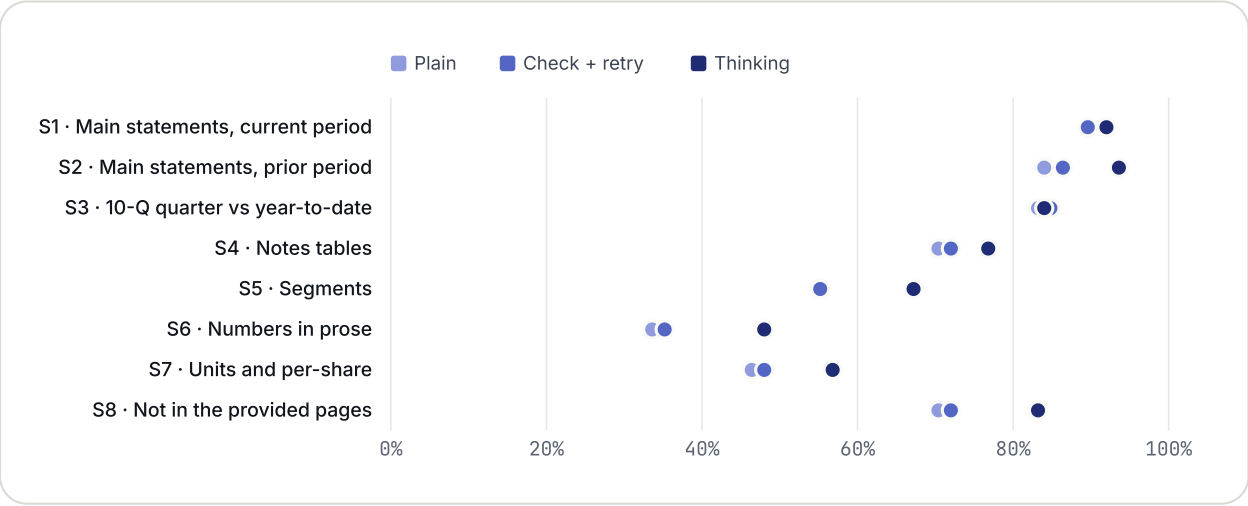

Figure 2. Accuracy per slice and setup (S8: correct "not found").

Plain setup: strongest on main statements, current period (89.6%) and main statements, prior period (84.0%); weakest on numbers in prose (33.6%) and units and per-share (46.4%).

Retry and thinking

Comparison	Accuracy change	Holm p	Confident-wrong change	Holm p
Check + retry vs plain	+1.3 pts	**0.00488**	+7	0.0938
Thinking vs plain	+8.0 pts	**1.14e-05**	-57	**3.73e-10**
Thinking vs check + retry	+6.7 pts	**0.000278**	-64	**3.17e-12**

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 196 answers again. Of those, 11 became correct, 7 became confident-wrong, 3 turned into "not found", and the rest still failed a check.

What stands out

- The model of the pilot study that motivated the benchmark, reported as a supplementary baseline outside the ten-model comparison (Qwen3.5-9B is the Alibaba entry).
- Weak at "not found" in the plain setup (88 of 125): asked for a quarter of an annual report, it answered "not found" only 6 times in 31.
- Thinking brings it to 74.1% and cuts confident-wrong answers from 94 to 37.

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`

UnitedHealth Group Incorporated · 10-Q · S3 10-Q quarter vs year-to-date · FE-0358

CORRECT

"What was UnitedHealth Group Incorporated's Net earnings attributable to UnitedHealth Group common shareholders for the three months ended September 30, 2026?"

Answer	`4110` in millions, page 3
Its quote	`Net earnings attributable to UnitedHealth Group common shareholders $ 4,110 $ 2,348 $ 15,874 $ 12,046`
Answer key	`4,110` million · from `Net earnings attributable to UnitedHealth Group common shareholders $ 4,110 $ 2,348 $ 15,874 $ 12,046`

Confident-wrong: passed every check, still wrong

Bank of America Corporation · 10-Q · S2 Main statements, prior period · FE-0141

CONFIDENT-WRONG

"What was Bank of America Corporation's other general operating (noninterest expense) for the nine months ended September 30, 2025?"

Answer	`912` in millions, page 48
Its quote	`Other general operating 2,333 912`
Answer key	`2,990` million · from `Other general operating 2,333 912 4,536 2,990`
Error type	Wrong period or column

Chevron Corporation · 10-Q · S4 Notes tables · FE-0404 CONFIDENT-WRONG

"What was Chevron Corporation's contributions from noncontrolling interests for the nine months ended September 30, 2026?"

Answer	`(417)` Millions of dollars, page 9
Its quote	`Net contributions from (distributions to) noncontrolling interests $ (417) $ (220)`
Answer key	`9` million · from `Contributions from noncontrolling interests 9 8`
Error type	Wrong item

MetLife, Inc. · 10-Q · S5 Segments · FE-0574 CONFIDENT-WRONG

"What was MetLife, Inc.'s effect of foreign currency translation and other - net for its EMEA segment for the nine months ended September 30, 2025?"

Answer	`342` in millions, page 42
Its quote	`Effect of foreign currency translation and other, net — — 342 219 149 3 713`
Answer key	`45` million · from `Effect of foreign currency translation and other, net — 32 102 45 — 179`
Error type	Wrong item

Right value, unusable evidence

JPMorgan Chase & Co. · 10-K · S1 Main statements, current period · FE-0048
RIGHT VALUE, UNUSABLE EVIDENCE

"What was JPMorgan Chase & Co.'s accrued interest and accounts receivable as of December 31, 2025?"

Answer	`111599` in millions, page 167
Its quote	`Accrued interest and accounts receivable 111,599 101,223`
Answer key	`111,599` million · from `Accrued interest and accounts receivable 111,599 101,223`
Checker	failed the span check
Error type	Wrong page cited

Wrong, caught by the checker

Home Depot, Inc. · 10-K · S4 Notes tables · FE-0414 CAUGHT WRONG

"What was Home Depot, Inc.'s total deferred tax liabilities as of February 1, 2026?"

Answer	`2553` in millions, page 65
Its quote	`Total deferred tax liabilities $ (2,553) $ (1,693)`
Answer key	`6,463` million · from `Total deferred tax liabilities (6,463) (5,481)`
Checker	failed the span check
Error type	Wrong item

How it was run

Hugging Face repository Qwen/Qwen3-8B (https://huggingface.co/Qwen/Qwen3-8B)	Pinned revision b968826d9c46
Licence Apache 2.0	Released (repository created) 2025-04-27
Precision bf16	Thinking switch {'chat_template_kwargs': {'enable_thinking': True}}

```
vllm serve Qwen/Qwen3-8B --served-model-name qwen3-8b --host 127.0.0.1 --
port 8000 --max-model-len 16384 --gpu-memory-utilization 0.9 --api-key
<key> --seed 0 --revision b968826d9c46dd6066d109eabc6255188de91218
```

```
uv run python hf_data.py pull
bash run_all.sh qwen3-8b                                  # serve, then C1, C2,
C3
uv run python runner.py qwen3-8b 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.