FILING-EVIDENCE · RESULTS PER MODEL

Qwen3.5-9B

Alibaba · 9.7B 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 It has **not been peer-reviewed** and has not been submitted to or accepted by any conference or journal.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

82.7%
accuracy with checkable evidence, plain setup (rank 1 of 10)

84.5%
best setup: check + retry

36
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	82.7% 80.1–85.1%	36 2.6–4.9%	124	83.0%	53	5.6 min
C2 · Check + retry	84.5% 81.9–86.7%	36 2.6–4.9%	124	84.7%	56	0.3 min
C3 · Thinking	74.6% 71.6–77.4%	16 1.0–2.6%	122	75.3%	3,574	99.2 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

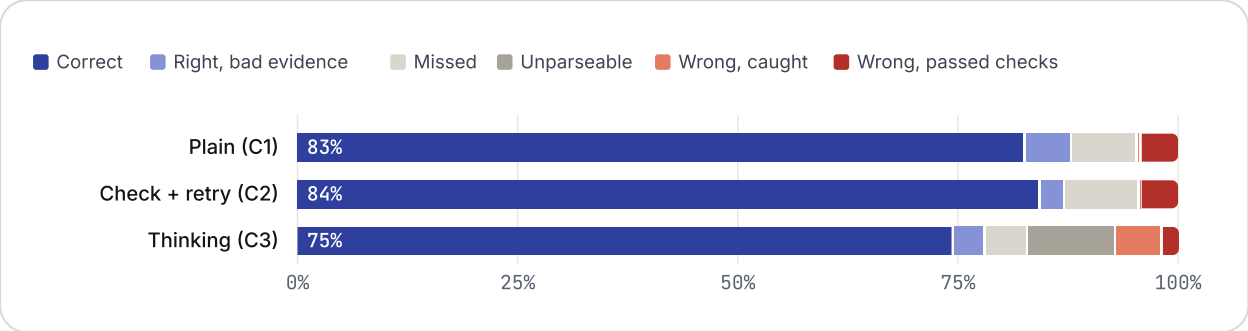

Figure 1. Outcomes of the 875 answerable questions in each setup.

In the plain setup, 724 answers were correct and 36 were wrong while passing every check: 2 scale or unit; 25 another period or column of the right line; 9 a different item. A further 46 had the right value with evidence the checker could not accept, and 4 wrong answers were rejected by the checker.

By kind of content

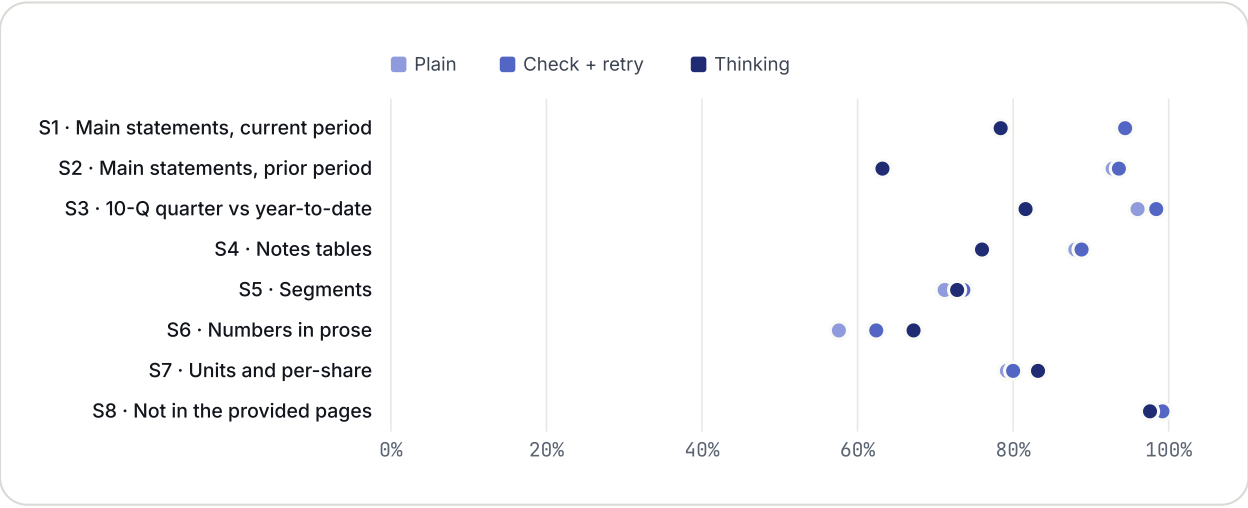

Figure 2. Accuracy per slice and setup (S8: correct "not found").

Plain setup: strongest on 10-Q quarter vs year-to-date (96.0%) and main statements, current period (94.4%); weakest on numbers in prose (57.6%) and segments (71.2%).

Retry and thinking

Comparison	Accuracy change	Holm p	Confident-wrong change	Holm p
Check + retry vs plain	+1.7 pts	0.000427	+0	1
Thinking vs plain	-8.1 pts	7.35e-05	-20	0.0107
Thinking vs check + retry	-9.8 pts	4.35e-07	-20	0.0107

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 50 answers again. Of those, 15 became correct, 0 became confident-wrong, 9 turned into "not found", and the rest still failed a check.

What stands out

- The most accurate model in the plain and check + retry setups, and significantly ahead of every other model in both.
- The fewest confident-wrong answers in every setup it was compared in; with thinking, 16 in 1,000, the lowest in the study.
- Thinking lowered its accuracy to 74.6% because it often reasoned past the pre-registered 8,192-token limit: 180 truncated answers, 88 of them unparseable. On the questions it finished, accuracy was 85.1%.

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-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)`

Apple Inc. · 10-K · S4 Notes tables · FE-0378 CORRECT

"What was Apple Inc.'s increases related to tax positions taken during a prior year for the fiscal year ended September 27, 2025?"

Answer	`1971` in millions, page 47
Its quote	`Increases related to tax positions taken during a prior year 1,267 1,971 1,727`
Answer key	`1,971` million · from `Increases related to tax positions taken during a prior year 1,267 1,971 1,727`

Confident-wrong: passed every check, still wrong

McDonald's Corporation · 10-Q · S2 Main statements, prior period · FE-0183 CONFIDENT-WRONG

"What was McDonald's Corporation's cash and equivalents increase (decrease) for the three months ended September 30, 2025?"

Answer	`371` in millions, page 5
Its quote	`Cash and equivalents increase (decrease) 371 537 419 1,328`
Answer key	`537` million · from `Cash and equivalents increase (decrease) 371 537 419 1,328`
Error type	Wrong period or column

MetLife, Inc. · 10-Q · S5 Segments · FE-0573 CONFIDENT-WRONG

"What was MetLife, Inc.'s effect of foreign currency translation and other - net for its asia 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	`32` million · from `Effect of foreign currency translation and other, net — 32 102 45 — 179`
Error type	Wrong item

Abbott Laboratories · 10-Q · S7 Units and per-share · FE-0756 CONFIDENT-WRONG

"What was Abbott Laboratories's Basic Earnings Per Common Share (per share) for the nine months ended September 30, 2025?"

Answer	`1.67` none, page 2
Its quote	`Basic Earnings Per Common Share $ 0.53 $ 0.94 $ 1.67 $ 2.72`
Answer key	`2.72` · from `Basic Earnings Per Common Share $ 0.53 $ 0.94 $ 1.67 $ 2.72`
Error type	Wrong period or column

Right value, unusable evidence

Progressive Corporation · 10-Q · S1 Main statements, current period · FE-0089
RIGHT VALUE, UNUSABLE EVIDENCE

"What was Progressive Corporation's total liabilities and shareholders' equity as of September 30, 2026?"

Answer	`127744` in millions, page 2
Its quote	`Total liabilities and shareholders' equity $ 127,744 $ 121,535 $ 123,039`
Answer key	`127,744` million · from `Total liabilities and shareholders' equity $ 127,744 $ 121,535 $ 123,039`
Checker	failed the span check
Error type	Quote not verbatim

How it was run

Hugging Face repository
Qwen/Qwen3.5-9B
(https://huggingface.co/Qwen/Qwen3.5-9B)

Pinned revision
`c20223623576`

Licence
Apache 2.0

Released (repository created)
2026-02-27

Precision
bf16

Thinking switch
`{'chat_template_kwargs': {'enable_thinking': True}}`

```
vllm serve Qwen/Qwen3.5-9B --served-model-name qwen3.5-9b --host 127.0.0.1
--port 8000 --max-model-len 16384 --gpu-memory-utilization 0.9 --api-key
<key> --seed 0 --revision c202236235762e1c871ad0ccb60c8ee5ba337b9a --
language-model-only --max-num-seqs 64
```

```
uv run python hf_data.py pull
bash run_all.sh qwen3.5-9b                                  # serve, then C1,
C2, C3
uv run python runner.py qwen3.5-9b 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.