FILING-EVIDENCE · RESULTS PER MODEL

GLM-4.6V-Flash

Zhipu (Z.ai) · 10.3B 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 It has **not been peer-reviewed** and has not been submitted to or accepted by any conference or journal.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

70.9%
accuracy with checkable evidence, plain setup (rank 4 of 10)

82.3%
best setup: thinking

110
wrong answers per 1,000 that passed every check (plain)

118/125
correct "not found" when the answer wasn't on the pages

Scorecard

Setup	Accuracy	Conf.-wrong	Abstention	Ign. scale	Tokens	Time
C1 · Plain	70.9% 67.8–73.8%	110 9.2–13.1%	118	78.5%	50	2.9 min
C2 · Check + retry	73.0% 70.0–75.9%	111 9.3–13.2%	118	80.7%	58	0.8 min
C3 · Thinking	82.3% 79.6–84.7%	61 4.8–7.8%	120	86.5%	385	14.0 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

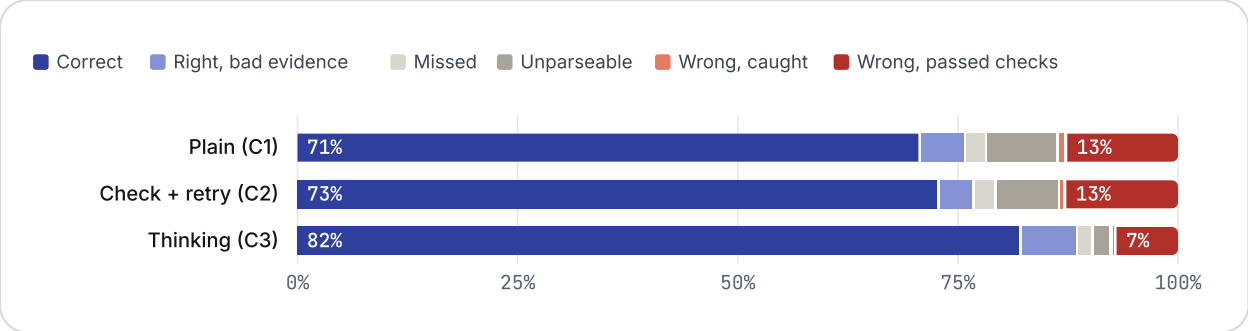

Figure 1. Outcomes of the 875 answerable questions in each setup.

In the plain setup, 620 answers were correct and 110 were wrong while passing every check: 67 scale or unit; 35 another period or column of the right line; 8 a different item. A further 45 had the right value with evidence the checker could not accept, and 8 wrong answers were rejected by the checker.

By kind of content

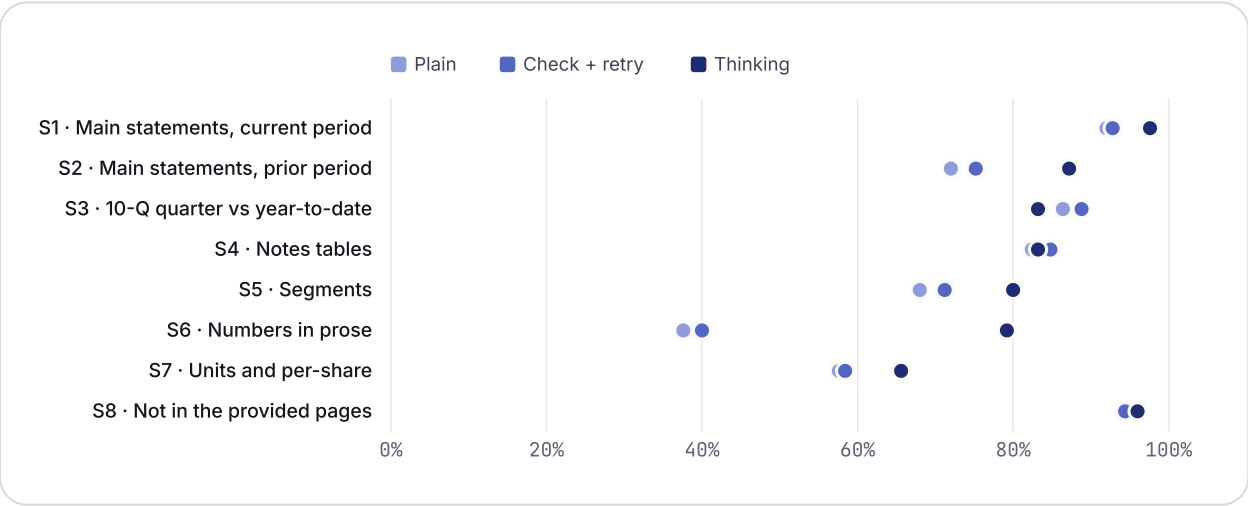

Figure 2. Accuracy per slice and setup (S8: correct "not found").

Plain setup: strongest on main statements, current period (92.0%) and 10-Q quarter vs year-to-date (86.4%); weakest on numbers in prose (37.6%) and units and per-share (57.6%).

Retry and thinking

Comparison	Accuracy change	Holm p	Confident-wrong change	Holm p
Check + retry vs plain	+2.2 pts	4.2e-05	+1	1
Thinking vs plain	+11.4 pts	1.45e-11	-49	2.62e-08
Thinking vs check + retry	+9.3 pts	4.08e-08	-50	1.6e-08

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 124 answers again. Of those, 19 became correct, 1 became confident-wrong, 1 turned into "not found", and the rest still failed a check.

What stands out

- The largest gain from thinking in the study (+11.4 points, to 82.3%, the highest thinking-mode accuracy), with short reasoning: 385 tokens per answer on average.
- Thinking also cut its confident-wrong answers from 110 to 61.
- A vision-language model, run text-only (the vision encoder was not loaded).

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)`

Caterpillar Inc. · 10-Q · S4 Notes tables · FE-0396 CORRECT

"What was Caterpillar Inc.'s stock-based compensation expense for the three months ended September 30, 2025?"

Answer	`66` in millions, page 35
Its quote	`Stock-based compensation expense (70) (66) (202) (196)`
Answer key	`66` million · from `Stock-based compensation expense (70) (66) (202) (196)`

Confident-wrong: passed every check, still wrong

Eli Lilly and Company · 10-K · S2 Main statements, prior period · FE-0177
CONFIDENT-WRONG

"What was Eli Lilly and Company's purchases of common stock for the fiscal year ended December 31, 2024?"

Answer	`4108` in millions, page 60
Its quote	`Purchases of common stock (4,108) (2,500) (750)`
Answer key	`2,500` million · from `Purchases of common stock (4,108) (2,500) (750)`
Error type	Wrong period or column

Wells Fargo & Company · 10-Q · S6 Numbers in prose · FE-0748 CONFIDENT-WRONG

"What was Wells Fargo & Company's net derivative gains (losses) from economic hedges related to mortgage loans held for sale and derivative loan commitments for the three months ended September 30, 2026?"

Answer	`-390` in millions, page 83
Its quote	`Net derivative gains (losses) from economic hedges (4) (390) 2 (471) 128`
Answer key	`87` million · from `(6)Includes net gains (losses) of $87 million and $116 million in the third quarter and first nine months of 2026, respectively, and $(6) million and $(20) million in the third quarter and first nine months of 2025, respectively, related to derivatives used as economic hedges of mortgage loans held for sale and derivative loan commitments.`
Error type	Wrong item

Oracle Corporation · 10-Q · S7 Units and per-share · FE-0828 CONFIDENT-WRONG

"What was Oracle Corporation's stock awards and shares excluded from calculation (number) for the three months ended August 31, 2026?"

Answer	`64` none, page 20
Its quote	`Stock awards and shares excluded from calculation(1) 64 2`
Answer key	`64` million · from `Stock awards and shares excluded from calculation(1) 64 2`
Error type	Scale or unit

Right value, unusable evidence

Progressive Corporation · 10-Q · S1 Main statements, current period · FE-0088
RIGHT VALUE, UNUSABLE EVIDENCE

"What was Progressive Corporation's net unrealized gains (losses) on fixed-maturity securities as of September 30, 2026?"

Answer	`2730` in millions, page 3
Its quote	`Net unrealized gains (losses) on fixed-maturity securities (2,730) 203 117`
Answer key	`2,730` million · from `Net unrealized gains (losses) on fixed-maturity securities (2,730) 203 117`
Checker	failed the span check
Error type	Wrong page cited

Wrong, caught by the checker

Walt Disney Company · 10-K · S7 Units and per-share · FE-0780 CAUGHT WRONG

"What was Walt Disney Company's tax impact of equity awards (%) for the fiscal year ended September 30, 2023?"

Answer	`1.1` %, page 101
Its quote	`Tax impact of equity awards (0.9) (1.1) (1.1)`
Answer key	`2.1` · from `Tax impact of equity awards (0.3) 0.8 2.1`
Checker	failed the span check
Error type	Wrong item

How it was run

Hugging Face repository zai-org/GLM-4.6V-Flash (https://huggingface.co/zai-org/GLM-4.6V-Flash)	Pinned revision 411bb4d77144
Licence MIT	Released (repository created) 2025-12-07
Precision bf16	Thinking switch {'chat_template_kwargs': {'enable_thinking': True}}

```
vllm serve zai-org/GLM-4.6V-Flash --served-model-name glm-4.6v-flash --host
127.0.0.1 --port 8000 --max-model-len 16384 --gpu-memory-utilization 0.9 --
api-key <key> --seed 0 --revision 411bb4d77144a3f03accbf4b780f5acb8b7cde4e
--language-model-only
```

```
uv run python hf_data.py pull
bash run_all.sh glm-4.6v-flash                                  # serve, then
C1, C2, C3
uv run python runner.py glm-4.6v-flash 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.