FILING-EVIDENCE · RESULTS PER MODEL

Phi-4

Microsoft · 14.7B parameters · fp8

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 It has **not been peer-reviewed** and has not been submitted to or accepted by any conference or journal.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

57.3%
accuracy with checkable evidence, plain setup (rank 8 of 10)

60.8%
best setup: check + retry

151
wrong answers per 1,000 that passed every check (plain)

116/125
correct "not found" when the answer wasn't on the pages

Scorecard

Setup	Accuracy	Conf.-wrong	Abstention	Ign. scale	Tokens	Time
C1 · Plain	57.3% 54.0–60.5%	151 13.0–17.5%	116	70.5%	59	6.3 min
C2 · Check + retry	60.8% 57.5–64.0%	167 14.5–19.1%	116	74.9%	75	1.6 min
C3 · Thinking	n/a: this model has no reasoning mode					

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

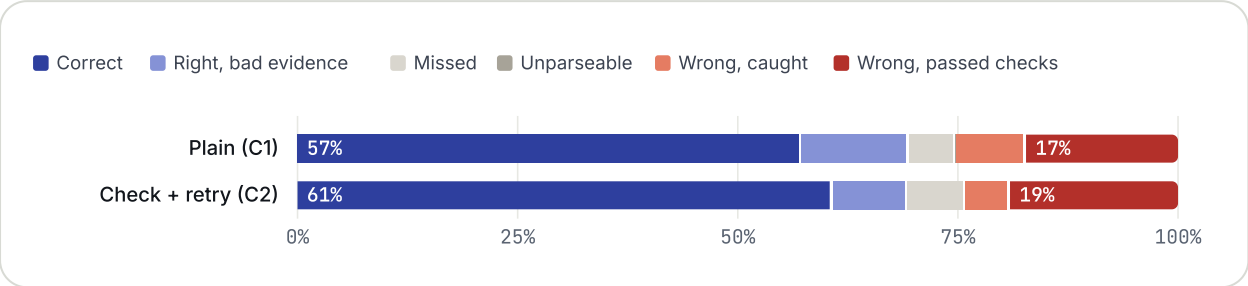

Figure 1. Outcomes of the 875 answerable questions in each setup.

In the plain setup, 501 answers were correct and 151 were wrong while passing every check: 116 scale or unit; 17 another period or column of the right line; 18 a different item. A further 107 had the right value with evidence the checker could not accept, and 69 wrong answers were rejected by the checker.

By kind of content

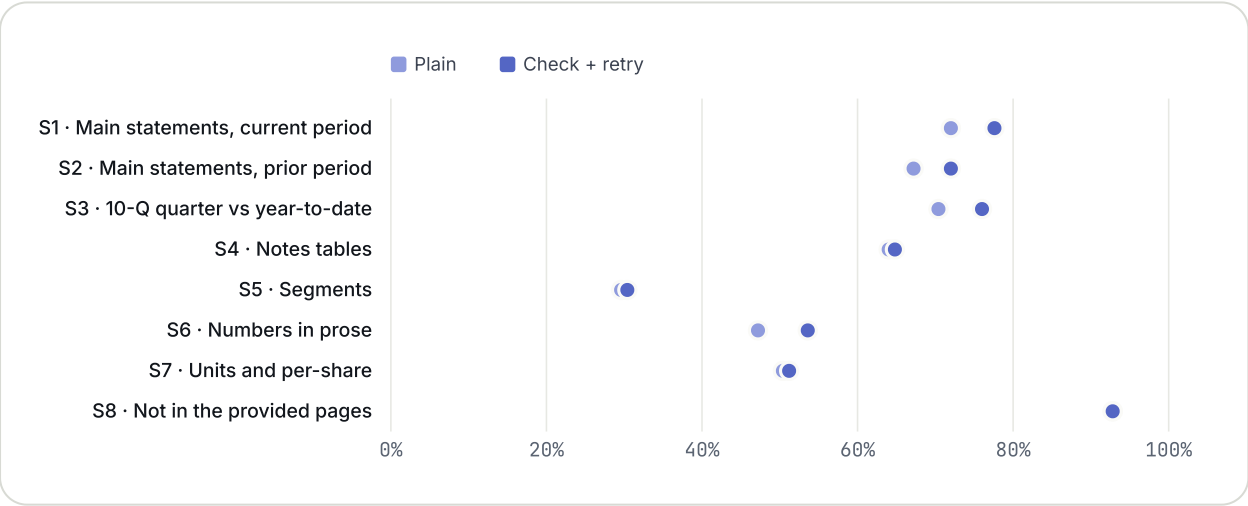

Figure 2. Accuracy per slice and setup (S8: correct "not found").

Plain setup: strongest on main statements, current period (72.0%) and 10-Q quarter vs year-to-date (70.4%); weakest on segments (29.6%) and numbers in prose (47.2%).

Retry and thinking

Comparison	Accuracy change	Holm p	Confident-wrong change	Holm p
Check + retry vs plain	+3.5 pts	1.96e-08	+16	0.000458

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 177 answers again. Of those, 31 became correct, 16 became confident-wrong, 12 turned into "not found", and the rest still failed a check.

What stands out

- The most confident-wrong answers of the ten (151 in the plain setup), and 116 of them are scale or unit errors: right line, right digits, but "scale: none" for a table in millions. Ignoring the stated scale its accuracy would be 70.5%.
- Retry makes this worse: 16 answers went from caught to confident-wrong after the retry.
- Run with FP8 weights (14.7B parameters do not fit in 24 GB in bf16).

Real answers

Chosen by a fixed rule (evenly spaced over question ids) from its best setup (check + retry), not picked by hand.

Correct

Adobe Inc. · 10-Q · S1 Main statements, current period · FE-0007 CORRECT

“What was Adobe Inc.'s cash and cash equivalents at end of period as of August 28, 2026?”

Answer	`4359` in millions, page 7
Its quote	`Cash and cash equivalents at end of period $ 4,359`
Answer key	`4,359` million · from `Cash and cash equivalents at end of period $ 4,359 $ 4,982`

Travelers Companies, Inc. · 10-Q · S3 10-Q quarter vs year-to-date · FE-0355 CORRECT

"What was Travelers Companies, Inc.'s income before income taxes for the nine months ended September 30, 2026?"

Answer	`7255` in millions, page 2
Its quote	`Income before income taxes 2,369 2,344 7,255 4,693`
Answer key	`7,255` million · from `Income before income taxes 2,369 2,344 7,255 4,693`

Confident-wrong: passed every check, still wrong

Chevron Corporation · 10-Q · S1 Main statements, current period · FE-0029
CONFIDENT-WRONG

"What was Chevron Corporation's deferred compensation and benefit plan trust as of September 30, 2026?"

Answer	`(240)` none, page 5
Its quote	`Deferred compensation and benefit plan trust (240) (240)`
Answer key	`240` million · from `Deferred compensation and benefit plan trust (240) (240)`
Error type	Scale or unit

Allstate Corporation · 10-Q · S4 Notes tables · FE-0386 CONFIDENT-WRONG

"What was Allstate Corporation's operating cash flow provided (used) for the nine months ended September 30, 2025?"

Answer	`183` in millions, page 43
Its quote	`Operating cash flow (used) provided (209) 183`
Answer key	`182` million · from `Operating cash flow provided (used) $ 209 $ (182)`
Error type	Wrong item

American Express Company · 10-Q · S6 Numbers in prose · FE-0637 CONFIDENT-WRONG

"What was American Express Company's net increase (decrease) in interest expense on long term debt and other for the three months ended September 30, 2025?"

Answer	`-488` in millions, page 64
Its quote	`Fixed-rate long-term debt $ 937 $ (89) $ 1,529 $ (488)`
Answer key	`47` million · from `We recognized in Interest expense on Long-term debt a decrease of $8 million and an increase of $47 million for the three months ended September 30, 2026 and 2025, respectively, and a decrease of $39 million and an increase of $88 million for the nine months ended September 30, 2026 and 2025, respectively, primarily related to the net settlements including interest accruals on our interest rate derivatives designated as fair value hedges.`
Error type	Wrong item

Right value, unusable evidence

General Dynamics Corporation · 10-Q · S1 Main statements, current period · FE-0037
RIGHT VALUE, UNUSABLE EVIDENCE

"What was General Dynamics Corporation's total current assets as of October 4, 2026?"

Answer	`26873` in millions, page 5
Its quote	`Total current assets $ 26,873 $ 24,248`
Answer key	`26,873` million · from `Total current assets 26,873 24,248`
Checker	failed the span check
Error type	Quote not verbatim

Wrong, caught by the checker

Charles Schwab Corporation · 10-Q · S1 Main statements, current period · FE-0097

CAUGHT WRONG

"What was Charles Schwab Corporation's Cash and Cash Equivalents, including Amounts Restricted at End of Period as of September 30, 2026?"

Answer	`68565` in millions, page 33
Its quote	`Cash and Cash Equivalents, including Amounts Restricted at End of Period $ 68,527 $ 54,865`
Answer key	`68,527` million · from `Cash and Cash Equivalents, including Amounts Restricted at End of Period $ 68,527 $ 54,865`
Checker	failed the value check
Error type	Wrong item

How it was run

Hugging Face repository microsoft/phi-4 (https://huggingface.co/microsoft/phi-4)	Pinned revision `2db69c1c3e91`
Licence MIT	Released (repository created) 2024-12-11
Precision fp8	Thinking switch `none (C3 n/a)`

```
vllm serve microsoft/phi-4 --served-model-name phi-4 --host 127.0.0.1 --
port 8000 --max-model-len 16384 --gpu-memory-utilization 0.9 --api-key
<key> --seed 0 --revision 2db69c1c3e91a05d2c64a3185acfbaf36f744e25 --
quantization fp8
```

```
uv run python hf_data.py pull
bash run_all.sh phi-4                                  # serve, then C1, C2
uv run python runner.py phi-4 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.