COMMUNITY RESEARCH REPORT · NOT PEER-REVIEWED · V1.0

Units, Columns and Periods: Presentation, Not Sector, Drives Errors When Small Open Models Read SEC Filings

Filing-Evidence Sectors: a pre-registered follow-up across ten sectors, 10,000 questions and eleven open models

Krishna Nandula · Pramana Research 23 November 2026 Version 1.0

Status: an independent community research report, published by Pramana Research on 23 November 2026. It has **not been peer-reviewed** and has not been submitted to or accepted by any conference or journal. It extends Filing-Evidence (arXiv:2611.24990) with the same method. Corrections and feedback are welcome on GitHub.

In brief

We asked eleven open AI models, each small enough to run on one gaming graphics card, to find 10,000 numbers in the SEC filings of 100 US companies across ten industries. Every answer had to quote the exact line it was read from, so a computer could check it. We wanted to know which numbers are hard, and why.

9 pts
accuracy gap between the easiest and hardest sector, against 67 points between the best and worst model

×2.2
odds of a quietly wrong answer when the unit ("in millions") is printed away from the number

88%
of quietly wrong answers on numbers with no unit on the page had the right digits at the wrong scale

4 of 5
models made significantly fewer quietly wrong answers when allowed to think; retrying added them for 8 of 10

1. **The industry hardly matters; the model matters most.** Pooled over models, accuracy runs from 61.2% (oil & gas) to 70.2% (utilities), and the sector effect is not significant once companies are taken into account. Models run from 19.2% to 86.4%.
2. **How a number is printed drives the errors.** Units stated away from the figure, more figures on the line, the answer sitting in a later column, and year-to-date periods each make a number significantly harder, with the other factors held fixed, and each keeps its direction under a second prompt and a second table layout.
3. **The one sector that looked dangerous wasn't, once we looked closely.** Oil and gas had the most quietly wrong answers (15.5%), but almost entirely because 14% of its numbers have no unit anywhere on the page. Without those, it is ordinary (10.1%).
4. **v1's main finding holds on ten times the data.** Thinking cuts quietly wrong answers (299 fixed, 108 introduced); telling a model its quote failed and letting it retry raises accuracy but adds quietly wrong answers.
5. **A quote check must survive formatting.** When tables were shown as Markdown, models made fewer column errors, but many dropped the table pipes when quoting, so a strict check rejected right answers. Any evidence checker needs to tolerate layout.

Abstract

Small open language models can extract numbers from SEC filings with evidence a machine can check, but they still make errors that pass the check. We ask which numbers cause those errors. In a pre-registered follow-up to Filing-Evidence, we built 10,000 questions over 200 recent 10-K and 10-Q filings of 100 US companies in ten sectors, with gold answers from the filers' inline XBRL and difficulty tags on every question, fixed before any model ran. Eleven instruct models (7–21B parameters) answered on one RTX 4090 in three setups: plain, check-and-retry, and thinking. Sector explains little: pooled accuracy ranges from 61.2% to 70.2% across sectors, against 19.2% to 86.4% across models, and the sector term is not significant ($p = 0.08$). Presentation explains much more: a unit stated away from the figure raises the odds of a wrong-but-passing answer 2.2-fold; each additional figure on the line, a later column and a year-to-date period each lower the odds of a correct answer (odds ratios 0.86, 0.69, 0.72). The worst sector for wrong-but-passing answers, oil and gas, is explained by numbers whose unit appears nowhere on the page; 88% of those errors are the right digits at the wrong scale. Thinking reduces such errors for 4 of 5 models; retrying after a failed check raises them for 8 of 10. The findings hold under a second prompt and a repeat run; Markdown tables cut wrong-column errors but expose how fragile a strict quote check is. We release the questions, every model answer, the code and the protocol.

1 · Introduction

Filing-Evidence asked whether small open models can read SEC filings with evidence a machine can check. They largely can: the best model answered 82.7% of its questions correctly with a verifiable quote. But a share of answers passed every automatic check and were still wrong: the right line, the wrong number. Those quiet errors are what make a model unsafe to trust.

v1 suggested where they come from: units printed in a header rather than next to the number, and neighbouring columns. With 1,000 questions it couldn't test that. This study can. We wrote the hypotheses down first, built 10,000 new questions across ten sectors with difficulty tags computed by code, froze everything, and only then ran the models. Three research questions, registered in advance:

- **RQ1.** How does accuracy with checkable evidence vary across ten industry sectors?
- **RQ2.** Which properties of a number make it hard: where its units are stated, how many figures share its line, which column it sits in, its period, and whether it is written in words?
- **RQ3.** Do v1's findings on check-and-retry and thinking hold on a larger and more varied set?

The answer to RQ1 is mostly no, and the answer to RQ2 explains why: what makes a number hard is how it is printed, and that varies within every sector more than between them.

2 · Design

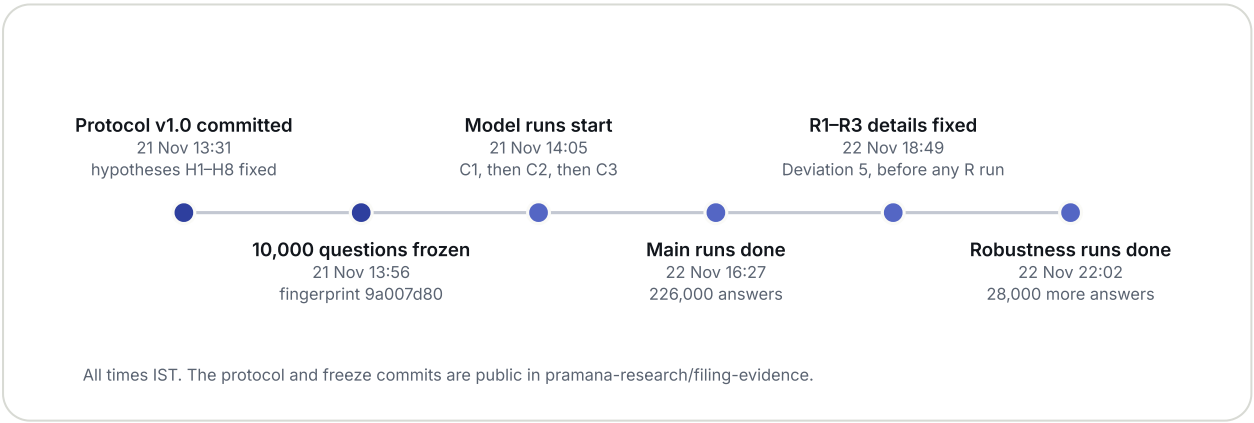

Figure 1. The study in time. The protocol and the frozen questions were committed publicly before any model ran; the robustness checks were fully specified before any of them ran.

Companies and filings. Ten sectors with ten companies each: banks, insurance, real estate, utilities, oil and gas, metals and mining, healthcare, technology, media and telecom. These were chosen because their statements are laid out differently. For each company we took its most recent 10-K and 10-Q, 200 filings in all. No company or datapoint from v1 is reused.

Questions and answer key. As in v1, gold answers come from the filers' own inline XBRL tags, and every question shows the model three pages: the answer page, a neighbouring page and a distractor. There are 1,000 questions per sector; 9,002 have an answer on the pages shown, and the rest test whether the model says "not found". The page rendering, wording rules, prompt and checker are v1's, unchanged.

Difficulty tags. Every question carries tags computed by code before any run: where the unit is stated (in the row label, in the line, in a table or column header, in a page header, or nowhere on the page), how many figures are on the evidence line, whether the answer is the first figure or a later one, the period length, and whether the amount is written with a word scale ("$12.6 billion").

Outcomes. As v1: *correct* (right value, the quote checks out); *confident-wrong* (wrong value, but the quote passes every check: the dangerous case); *caught-wrong* (wrong, and the check fails); *unverifiable-right* (right value, unusable quote); *missed*; and for absent numbers *abstained correctly* or *invented*.

3 · Models and setups

The ten models of v1 at the same pinned revisions, plus Qwen3-8B as a supplementary model, all served with vLLM 0.33 on one RTX 4090 capped at 450 W, temperature 0. Three setups: **plain** (C1) and **check + retry** (C2) on all 10,000 questions, and **thinking** (C3) on a stratified 1,000-question subset for the six models with a reasoning mode, with up to 8,192 new tokens. In total 251,000 answers, including the robustness checks of section 7.

4 · Does the sector matter? (RQ1, H1)

	Banks	Utilities	Health	Insurance	Media	Metals	Oil & gas	Real est.	Tech	Telecom	All
Qwen3.5-9B	87	90	89	82	86	88	85	89	86	82	**86**
gpt-oss-20b	80	85	80	77	78	75	75	80	80	76	**79**
GLM-4.6V-Flash	77	83	77	76	77	77	73	76	77	75	**77**
Granite-4.2-8B	76	84	78	72	74	75	73	73	78	75	**76**
Ministral-3-8B	72	79	74	72	73	69	66	71	73	71	**72**
Qwen3-8B *	69	77	68	71	70	70	65	76	71	71	**71**
Llama-3.1-8B	66	72	63	63	67	63	63	65	64	59	**64**
Gemma-3-12B	62	70	63	65	66	58	62	59	66	68	**64**
Phi-4	66	65	58	62	67	66	56	64	65	64	**63**
Nemotron-Nano-9B-v2	48	53	47	42	46	48	44	44	45	44	**46**
OLMo-3-7B	19	22	21	17	21	22	16	18	17	18	**19**

Accuracy (% of answerable), plain setup. * supplementary model.

Figure 2. Accuracy by model and sector, plain setup. Rows differ far more than columns.

Not much. Pooled over the ten main models, accuracy ranges from 61.2% (oil & gas) to 70.2% (utilities): a spread of 9 points, against 67 points between the best and worst model. The order of the models is much the same in every sector.

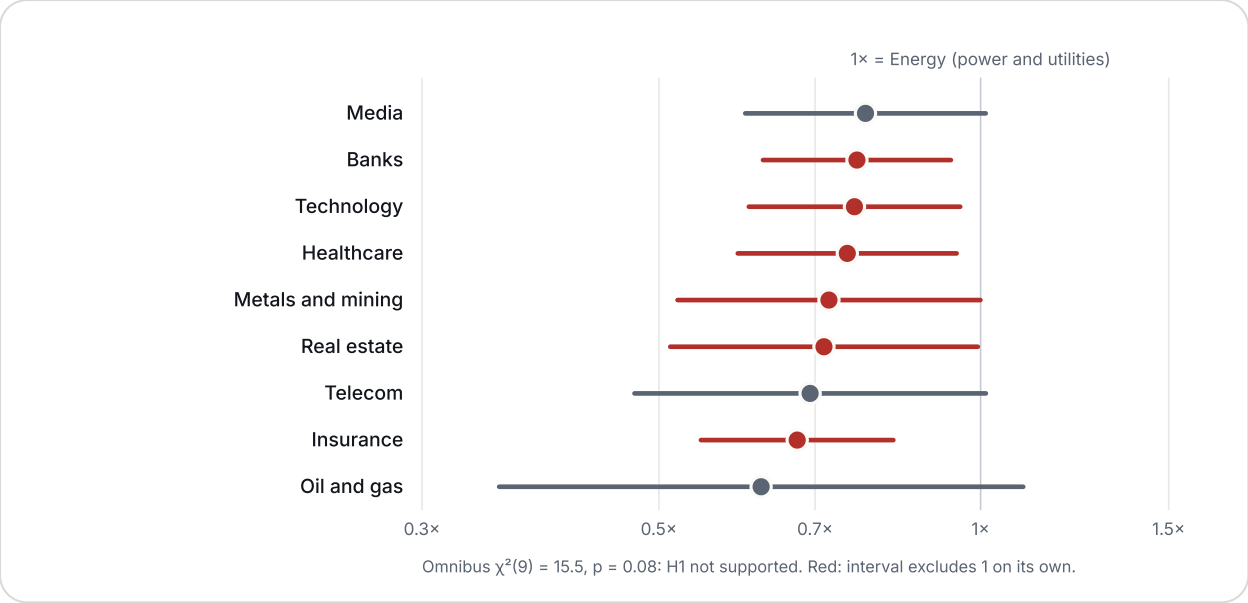

Figure 3. Odds of a correct answer in each sector compared with the easiest, from a logistic regression with a model fixed effect and standard errors clustered by company.

The pre-registered test (H1) is a logistic regression of correctness on sector with a model fixed effect and standard errors clustered by company: $\chi^2(9) = 15.5$, $p = 0.08$. **H1 is not supported.** Companies within a sector differ as much as sectors do. Insurance is the one sector whose interval excludes the easiest sector on its own (odds ratio 0.67), but H1 is judged on the omnibus test.

The exception that wasn't. Confident-wrong answers vary more by sector than accuracy does, and oil and gas stands out at 15.5% of answerable questions, about twice the rate of telecom or technology. The reason is specific and checkable: 14% of its questions have no unit stated anywhere on the pages shown (the scale appears only on another page of the filing), against 0–8% elsewhere. Those questions are answered confident-wrong 48% of the time. Without them, oil and gas falls to 10.1%, in line with every other sector. This analysis was not pre-registered and is reported as exploratory.

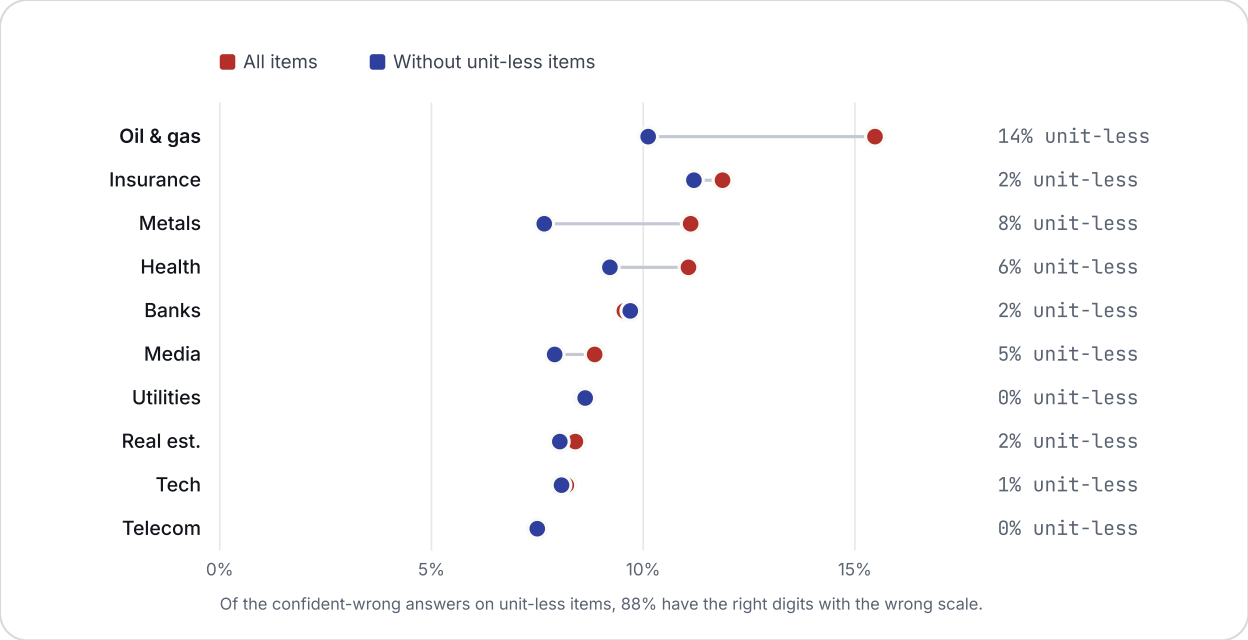

Figure 4. Confident-wrong rate by sector, with all items (red) and without items whose unit is stated nowhere on the pages (blue).

5 · What makes a number hard? (RQ2, H2–H6)

All five pre-registered hypotheses about the tags are supported. Each is a term in one logistic regression on all tags with a model fixed effect, standard errors clustered by company (89,780 answers, plain setup, ten main models), so each effect is estimated holding the others fixed.

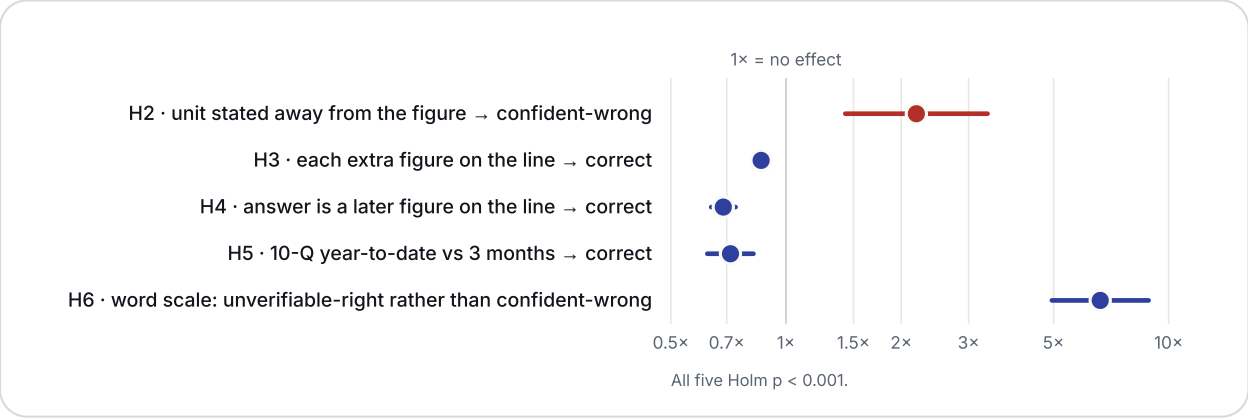

Figure 5. Odds ratios with 95% intervals for H2–H6. All are significant after Holm correction across H1–H8.

- **H2 · Units away from the figure.** When the unit is in a table or page header ("in thousands") rather than in the row label, the odds of a confident-wrong answer are 2.2 times higher (9.3% vs 4.4%). When the unit is nowhere on the page, the rate reaches 41%.
- **H3 · Crowded lines.** Each extra figure on the evidence line lowers the odds of a correct answer by a factor of 0.86: accuracy falls from 71% on two-figure lines to 42% on lines with six or more.

- **H4 · Later columns.** When the answer is not the first figure on its line, the odds fall by a factor of 0.69 (70% vs 60%).
- **H5 · Year-to-date figures.** In 10-Qs, six- and nine-month figures are harder than three-month figures from the same filings (odds ratio 0.72; 61% vs 69%).
- **H6 · Word scales.** Amounts written as "$12.6 billion" fail differently: mostly as unverifiable-right (the model restates the value in other units, so its quote no longer matches), not as confident-wrong (1,840 vs 354; odds ratio 6.6).

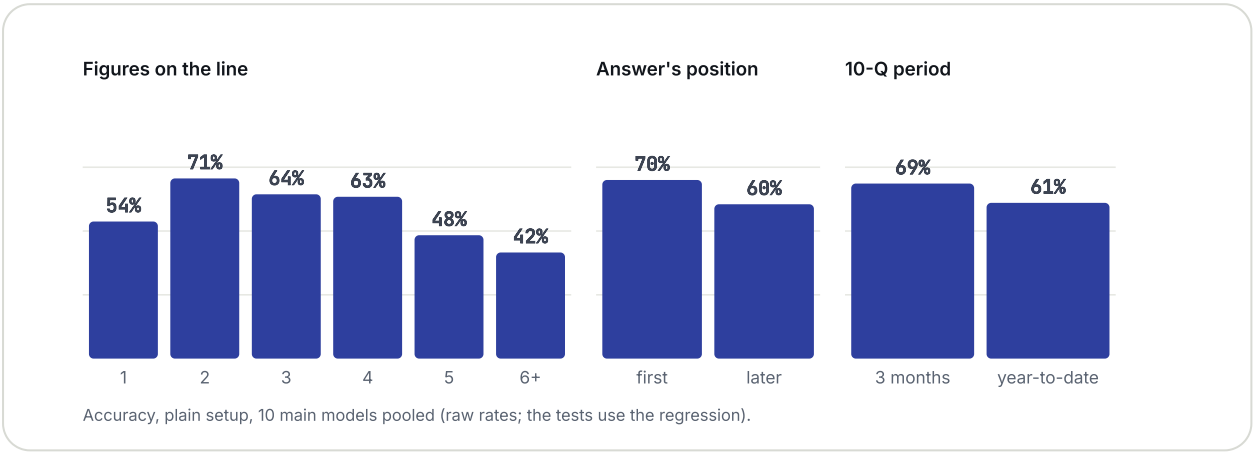

Figure 6. Raw accuracy by where the number sits. The tests use the regression above; these rates show the size of the effects.

The mechanism: scale guessing. Of the confident-wrong answers on numbers whose unit is stated nowhere on the page, 88% have exactly the right digits with the wrong scale: "52,611 in thousands" where the filing means millions. The quote passes because the digits match. A checker that compares digits cannot catch this; only knowing the unit can.

6 · Do v1's findings hold? (RQ3, H7)

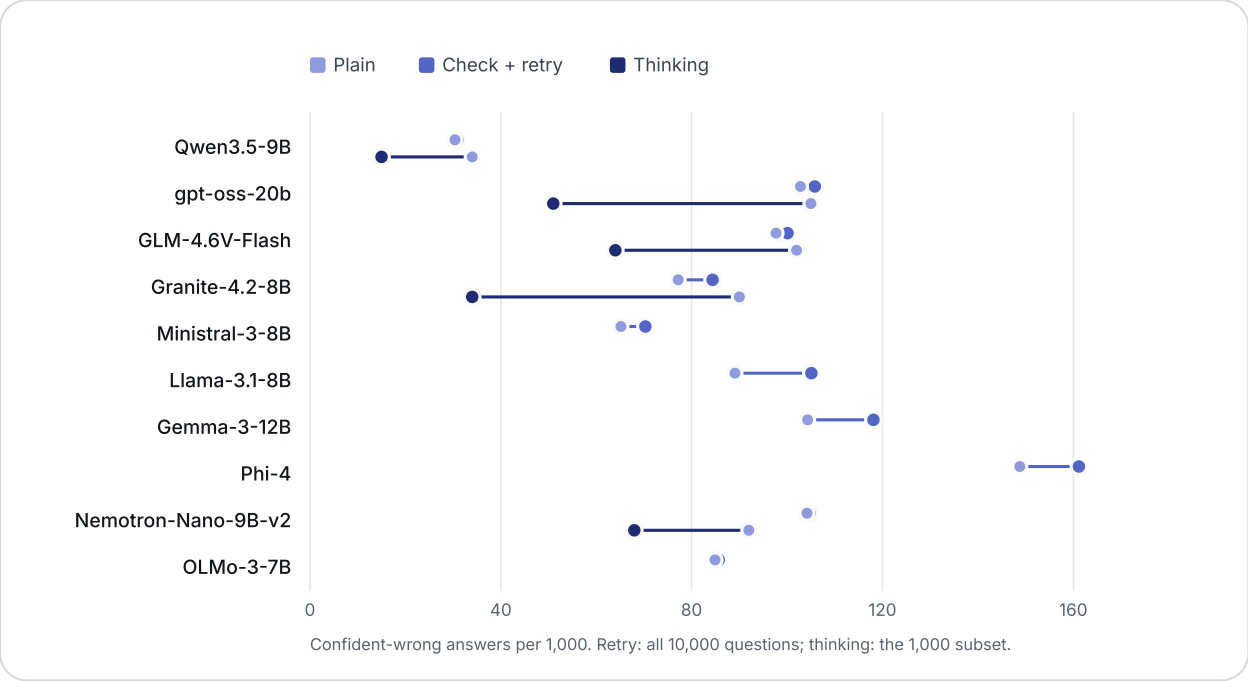

Figure 7. Confident-wrong answers per 1,000 in the plain setup and after check + retry (all questions) or thinking (the 1,000-question subset).

Yes. **Thinking lowers confident-wrong answers** significantly for 4 of the 5 main models with a reasoning mode (paired exact McNemar, Holm-corrected within model); pooled, 299 answers stop being confident-wrong and 108 start. **Check + retry does not**: it never lowers them, and raises them significantly for 8 of 10 models, while raising accuracy for 9. Told that its quote failed, a model often finds another line containing its wrong number. **H7 is supported.**

One exception needs stating plainly. Qwen3.5-9B's accuracy fell with thinking (743 → 689 of 900) because 204 of its 1,000 answers ran into the pre-registered 8,192-token limit and were cut off. Its confident-wrong answers still fell (34 → 15). A larger budget might change its accuracy result; we report the limit as fixed.

7 · How robust is this?

Three checks, fixed in the protocol and specified in full before any of them ran, test whether the findings depend on how we asked rather than what we asked. All run in the plain setup on the 1,000-question subset.

R3 · A repeat run. Running the identical configuration again changes the outcome of 0.6%–3.4% of questions for the three models tested, with no significant difference in accuracy or confident-wrong. Single runs are stable.

R2 · A different prompt. A second prompt with different wording and a different worked example (a 10-Q with quarter and year-to-date columns) leaves the ranking of models almost unchanged (Kendall's $\tau = 0.87$). Most models move by a few points. The two weakest models

are the exception: Nemotron-Nano-9B-v2 and OLMo-3-7B lose accuracy and gain confident-wrong answers under the new wording.

R1 · Markdown tables (H8). The same pages, with tables shown as Markdown: one row per line, column headers first. Wrong-period-or-column errors fall from 349 to 286 (pooled exact McNemar, p = < 0.001). **H8 is supported.** Confident-wrong answers fall for every model. But strict accuracy collapses for most models, because many quote the row without its pipes ("Total deposits 239,486" instead of "| Total deposits | 239,486 |") and the pre-registered check, applied to the page text exactly as shown, rejects the quote. Under that check the model ranking scrambles (τ = 0.20).

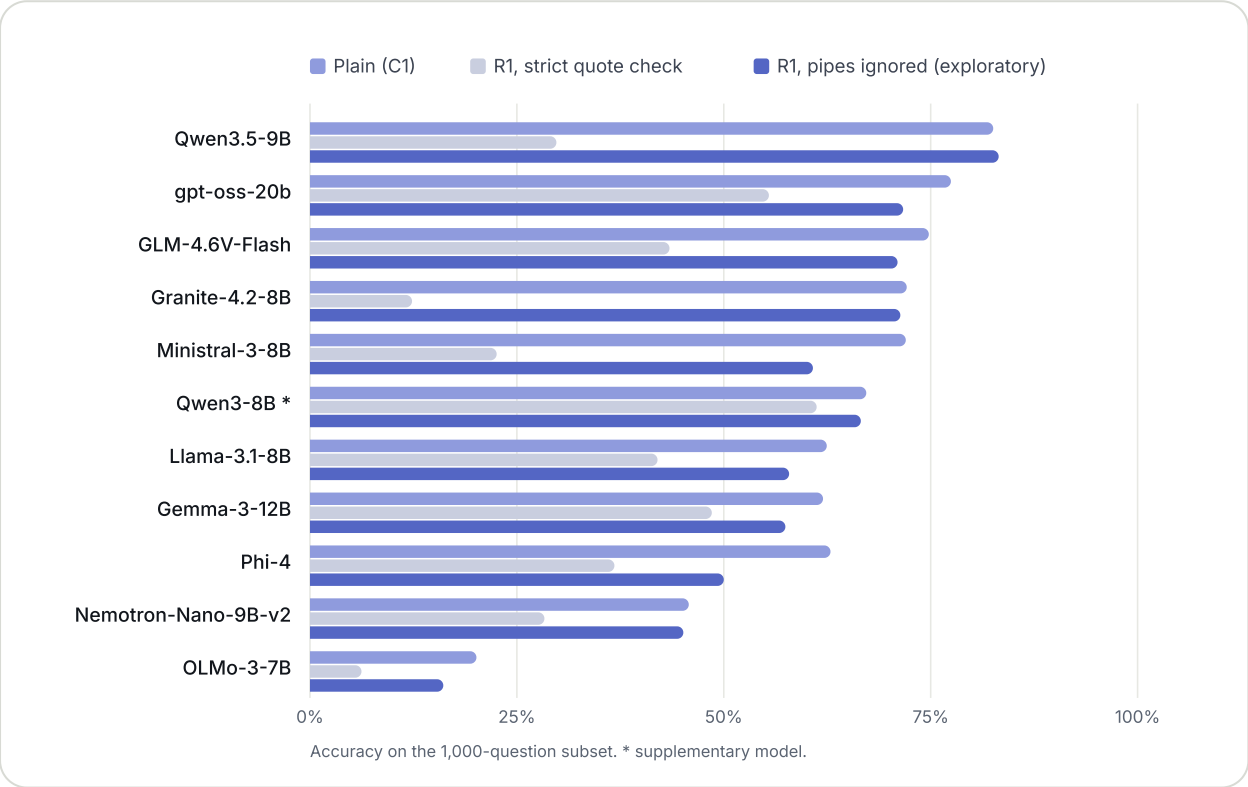

Figure 8. Accuracy in the plain setup, with Markdown tables under the strict quote check, and with Markdown tables when the pipes are ignored (exploratory).

To separate quoting from reading we graded R1 a second time with the pipes ignored, a decision made after seeing the strict results and therefore exploratory. Most of the collapse disappears, and the ranking returns (τ = 0.87). The best models read Markdown tables as well as flat text; the others lose up to 13 points (Phi-4). Both gradings tell the same practical story: layout changes which errors a model makes, and an evidence check must be robust to it.

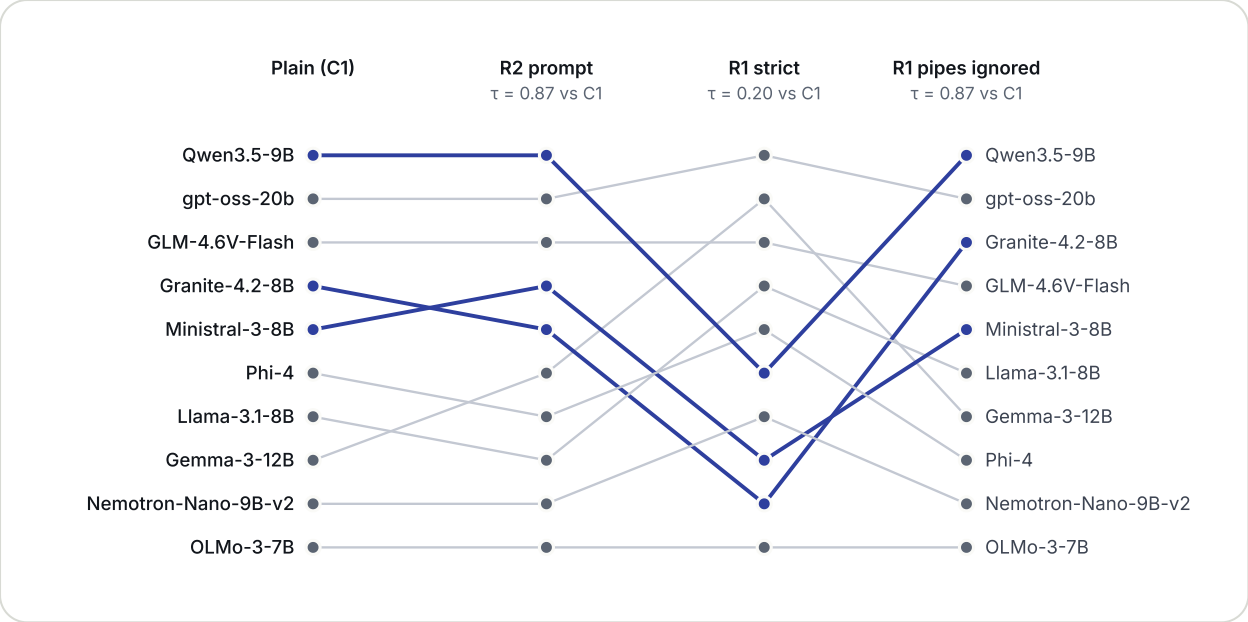

Figure 9. Model ranks by accuracy under each check.

The tag effects of H2–H4 keep their direction under R1 and R2. Together with the stable repeat run, the headline findings are **robust** in the sense the protocol defines.

8 · Discussion

Look at the line, not the industry. Anyone deploying a model to read filings should worry less about the sector and more about how each number is printed. Numbers whose unit sits in a header, crowded lines, prior-period columns and year-to-date figures are where errors concentrate, in every sector.

Give the model the unit. The most dangerous single case is a number whose unit isn't on the page at all. The model guesses, and a digit-based check passes the guess. A reader system should carry the document's scale statement alongside every page it shows, or have the checker verify the scale separately.

Think, don't retry. Thinking reduces quiet errors; retrying after a failed check increases them. If budget allows one intervention, it should be thinking, with a token limit large enough not to cut answers off.

Make the evidence check layout-tolerant. A strict character-for-character check is easy to explain but punishes models for formatting, not reading. Comparing the words and figures of a quote, not its punctuation, keeps the guarantee and removes the artefact.

9 · Limitations

- **Associations, not interventions.** The tags describe real filings; we did not move a unit from a row label to a header and re-ask. The regression holds the other tags fixed and R1 is a true intervention on layout, but the tag effects are observational.

- **One hardware and serving configuration.** One RTX 4090 at 450 W, vLLM 0.33, temperature 0.
- **A fixed thinking budget.** 8,192 new tokens cut off a fifth of Qwen3.5-9B's thinking answers.
- **R1's rendering is imperfect.** SEC tables are irregular; in 75% of tables with a header, the header row has as many cells as the first data row. The rest are misaligned to some degree.
- **Exploratory parts are labelled.** The oil-and-gas analysis and the pipe-insensitive R1 grading were not pre-registered as tests.
- **US filings only,** in English, from companies large enough to file inline XBRL.

10 · Reproducibility

Everything is public: the protocol (every change after it was finalised is logged in its §14, with the date and whether any result existed at the time), the code, the frozen questions (fingerprint `9a007d80`), the Markdown pages of R1 (fingerprint `7cdf066e`) and every model reply with its reasoning. In a fresh clone, `python hf_data.py pull` downloads the data and verifies both fingerprints; re-grading every stored reply reproduces every `summary.json` byte for byte, and `analyze.py` reproduces every statistic in this report. The pre-registration commits are public in pramana-research/filing-evidence (https://github.com/pramana-research/filing-evidence/commit/00260c43e78ad0f53fcc23ddec6645b97c2559b7).

Results per model

Each model has its own page with its accuracy and confident-wrong rate by sector and its results in every setup and check.

Qwen3.5-9B
main model
Plain 86.4% CW 3.0%

gpt-oss-20b
main model
Plain 78.6% CW 10.3%

GLM-4.6V-Flash
main model
Plain 76.9% CW 9.8%

Granite-4.2-8B
main model
Plain 75.9% CW 7.7%

Ministral-3-8B
main model
Plain 71.9% CW 6.5%

Llama-3.1-8B
main model
Plain 64.4% CW 8.9%

Gemma-3-12B

Phi-4

main model

Plain 63.9% CW 10.4%

main model

Plain 63.3% CW 14.9%

Nemotron-Nano-9B-v2

main model

Plain 45.9% CW 10.4%

OLMo-3-7B

main model

Plain 19.2% CW 8.5%

Qwen3-8B

supplementary

Plain 70.8% CW 9.4%

References

1. K. Nandula (2026). Show Your Work: Can Small Open Models Read SEC Filings with Evidence a Machine Can Check? Pramana Research. arXiv:2611.24990 (https://arxiv.org/abs/2611.24990).
2. U.S. Securities and Exchange Commission. EDGAR full-text and submissions APIs. sec.gov/edgar (https://www.sec.gov/edgar).
3. XBRL International. Inline XBRL Specification 1.1. xbrl.org (https://www.xbrl.org/specification/inlinexbrl-part1/rec-2013-11-18/inlinexbrl-part1-rec-2013-11-18.html).
4. E. B. Wilson (1927). Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association* 22(158), 209–212.
5. Q. McNemar (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika* 12(2), 153–157.
6. S. Holm (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics* 6(2), 65–70.
7. M. G. Kendall (1938). A new measure of rank correlation. *Biometrika* 30(1–2), 81–93.
8. W. Kwon et al. (2023). Efficient memory management for large language model serving with PagedAttention. *SOSP '23*. (vLLM)
9. Model cards of the eleven models evaluated, on Hugging Face, at the pinned revisions recorded with the results.

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.

How to cite

```
@misc{pramana2026filingevidencesectors,
  title        = {Units, Columns and Periods: Presentation, Not Sector, Drives
Errors When Small Open Models Read SEC Filings},
  author       = {Nandula, Krishna},
  organization = {Pramana Research},
  year         = {2026},
  month        = {nov},
  url          = {https://pramana-research.com/research/filing-evidence-
sectors},
  note         = {Community research report, not peer-reviewed. Filing-Evidence
Sectors, version 1.0}
}
```

This is a community research report; it has not been peer-reviewed. Text, data and figures are licensed under CC BY 4.0 and the code under the Apache License 2.0. Model outputs remain subject to each model's licence. Nothing here is investment advice.

© 2026 Pramana Research · Text, data and figures: CC BY 4.0 · Code: Apache 2.0 · Legal

Model outputs are not investment advice. Check every figure against its source.

Pramana Research is an independent research group, not affiliated with any other company of a similar name.