Filing-Evidence explained: can small open LLMs read SEC filings with evidence a machine can check?

24 November 2026 · 8 min read · blogs.iictp.org/filing-evidence-small-llms-sec-filings-inline-xbrl-explained

Short answer: Filing-Evidence is a benchmark from an arXiv preprint (submitted 18 November 2026) that tests whether small open language models can pull numbers from SEC 10-K and 10-Q filings and also give the page and a verbatim quote a machine can check. The authors report accuracy with checkable evidence between 17.1% and 82.7% for ten models on a single consumer GPU, and that between 16 and 167 wrong answers per 1,000 still pass every automatic check. This is an arXiv preprint. It has not been peer reviewed, and its numbers are the authors' own claims.

TL;DR The short version

- Filing-Evidence tests whether small open models can extract numbers from SEC filings and give a page and quote a machine can check.
- Accuracy with checkable evidence ranged from 17.1% to 82.7% across ten models on one consumer GPU.
- Between 16 and 167 wrong answers per 1,000 still passed every automatic check, so 'with a citation' is not the same as 'correct'.

Read online, listen to it, or change the text size: https://blogs.iictp.org/filing-evidence-small-llms-sec-filings-inline-xbrl-explained

What is Filing-Evidence?

The paper, "Show Your Work: Can Small Open Models Read SEC Filings with Evidence a Machine Can Check?", starts from the point that a fluent answer is not a verified one. Filing-Evidence is a set of 1,000 questions over 90 recent 10-K and 10-Q filings from 45 US companies. The answer keys are taken from the filers' own inline XBRL tags, so the correct value is machine-defined rather than hand-labelled.

How it works

- **Evidence requirement.** A model must return the number plus the page and a verbatim quote from the filing that contains it. An automatic checker verifies the evidence.
- **Ten models, three setups.** The authors tested ten instruct models from ten developers, 7 to 21 billion parameters, on one RTX 4090 GPU, in a plain setup, a check-and-retry setup and the model's own reasoning mode.
- **Ground truth from XBRL.** Because inline XBRL tags attach a defined concept, period and unit to each reported number, the key is unambiguous.

Results reported by the source

- Accuracy with checkable evidence ranged from 17.1% to 82.7% in the plain setup, and the best result, 84.5%, came from Qwen3.5-9B.
- The checker rejects many wrong answers, but between 16 and 167 answers per 1,000 give a wrong value for an answerable question and still pass every check.
- In the plain setup, 57% of those undetected errors are scale or unit errors, and 30% take another period or column of the right line.
- Reasoning mode significantly reduced such errors for 4 of the 5 models that offer it. Retrying after a failed check raised accuracy by up to 3.9 points and never lowered it.

Limitations and open questions

- It is a preprint with the authors' own protocol, and the paper says questions, outputs and code are public so others can reproduce it.
- The questions are about numbers stated in the filing. They do not test judgement, such as whether a disclosure is adequate.
- A quote that contains the number is evidence the number exists, not that it is the right period, unit or column, which is exactly where the surviving errors sit.

How to apply the lesson in finance teams

1. Never accept an extracted figure without its source page and quote, and check that the quote contains the figure.
2. Add checks for the errors the paper found: unit and scale (thousands versus millions) and period or column mismatch.
3. Where a filing is tagged in inline XBRL, compare against the tag's value, period and unit instead of relying on text alone.
4. Keep a human sign-off for any figure that goes into a model, a memo or a filing.

What it means for accountants and analysts

For review work, "with a citation" is not "correct". The same inline XBRL tags that regulators collect are a cheap source of ground truth for testing any document-reading tool. See the SEC roundtable on machine-understandable data and DNB on AI and data standards.

Frequently asked questions

- **What is Filing-Evidence?** A benchmark of 1,000 questions over 90 SEC filings from 45 US companies, with answer keys from inline XBRL tags.
- **How good were small models?** Between 17.1% and 82.7% accuracy with checkable evidence in the plain setup, and up to 84.5% in the best case, according to the authors.
- **What errors slipped through?** Mostly scale or unit mistakes and the wrong period or column on the right line.
- **Is it peer reviewed?** No, it is an arXiv preprint.

Related reading

- SEC roundtable on machine-understandable data and XBRL
- DNB, XBRL and Project TAVI
- What is LLM hallucination?
- AI Practitioner

Sources and currency

Sources were read on 24 November 2026. Products, rules and standards change, so check the linked originals before you rely on any detail.

> Related credential: AI Practitioner. Explore the exam facts and syllabus.

Sources and further reading

1. Krishna Nandula. Show Your Work: Can Small Open Models Read SEC Filings with Evidence a Machine Can Check? arXiv 2611.24990, 18 November 2026 (preprint). arxiv.org/abs/2611.24990.

Published by IICTP, an independent technical credential body, on 24 November 2026. The current version of this post is at https://blogs.iictp.org/filing-evidence-small-llms-sec-filings-inline-xbrl-explained. More guides: blogs.iictp.org.