XBRL calculation errors: why an LLM benchmark's error rate was mostly not arithmetic

27 November 2026 · 5 min read · blogs.iictp.org/xbrl-calculation-errors-llm-benchmark-dqc-rules

Short answer: a working paper by Vladimir Beskorovainyi, posted on arXiv on 20 November 2026, shows that the calculation-error rate for XBRL tasks in the FinReason Cup 2026 does not mainly count arithmetic mistakes. A rule-based system with no learned parameters scored over 95% on the test set. It matters because teams choosing AI for XBRL validation need to know what an error category really measures.

TL;DR The short version

- A working paper posted on 20 November 2026 shows the calculation-error rate for XBRL tasks in the FinReason Cup 2026 mostly does not count arithmetic mistakes.
- A rule-based system with no learned parameters scored over 95% on the test set.
- Before choosing AI for XBRL validation, check what each error category really measures and compare against deterministic rules.

Read online, listen to it, or change the text size: https://blogs.iictp.org/xbrl-calculation-errors-llm-benchmark-dqc-rules

What are XBRL calculation errors here?

XBRL is a structured format for financial reports in which each number is a tagged fact, and calculation relationships say which facts should add up to which. In the benchmark task, a system returns two values per case: the one the filing reports and the one the relationships imply. Failures are labelled structural, extraction or calculation.

How the analysis works

The cases come from three data quality rules of the XBRL US Data Quality Committee: DQC_US_0015, DQC_US_0117 and DQC_US_0126. The author rebuilt them as deterministic code. For one rule, DQC_US_0015, the correct answer is the negation of the reported value, so no sum remains to be done. Using published scores, the author bounds how many errors labelled calculation fall on such cases.

What is new compared with a leaderboard reading

A leaderboard suggests that the remaining weakness of strong systems is arithmetic. The paper argues that the label often records a wrong fact or a wrong route through the rules. Because scoring is hierarchical, accepting a reported value that is wrong can move a failure from the extraction column into the calculation column.

Results reported by the author

- The negation pattern held in all 110 published practice cases for DQC_US_0015, against 1 of 120 and 6 of 102 for the other two rules.
- The lower bound on non-arithmetic errors was positive for 22 of 23 ranked entries.
- For the two joint-first entries, at least 11 of 12 and 12 of 13 calculation errors fell on cases with no arithmetic left.
- The deterministic system ran all 680 cases in 1.4 seconds, scoring 96.62% on development data with the offline judge and 95.29% on the published test set.
- Remaining errors clustered where several facts were equally flagged and the question named no dimension member.

This is an arXiv preprint or workshop paper. Treat its numbers as the authors' own results, not as independently confirmed findings.

Limitations and open questions

- This is one participant's working notes on one competition task.
- The system targets three known rules and is not a general XBRL validator.
- The practice data had a different rule mix from the ranked data, so practice results did not carry over directly.
- The bounds rely on published scores, since the ranked answers were not released.

Sources were read on 27 November 2026. Products, rules and data change, so check the linked originals before you rely on any detail.

How to apply it to filing checks

1. Run deterministic validation first for any rule that can be written down.
2. Use a model for the parts that need judgment, such as choosing among ambiguous facts, and log the choice.
3. Break error reports down by rule, not only by error type.
4. Require the system to name the context, period and dimension of every fact it used.
5. Test on filings with several facts for the same concept.
6. Keep a human decision on any correction to a filed value.

What it means for reporting and audit teams

Ask vendors for per-rule results and the share of cases a plain rules engine would solve. A model that is expensive and slower than code for a coded rule adds risk without benefit. Background on structured evidence is in filing evidence with small models.

Frequently asked questions

- **What are XBRL calculation errors in this benchmark?** Failures labelled as calculation problems when the implied value differs from the expected one.
- **Were they mostly arithmetic?** The author shows that for strong entries they largely were not.
- **Did the rule system use AI?** No learned parameters, training step or external service.

- **Which rules were covered?** DQC_US_0015, DQC_US_0117 and DQC_US_0126.

Related reading

- Reading SEC filings with checkable evidence
- Auditable financial knowledge graphs
- FRC 2027 digital reporting taxonomies
- AI Practitioner certification

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

Sources and further reading

1. Vladimir Beskorovainyi. How Much of a Calculation-Error Rate Is Not Arithmetic: Auditing the FinReason Cup 2026 Task 3 Metric with a Deterministic Rule System. arXiv 2611.27192. 20 November 2026 (working notes). arxiv.org/abs/2611.27192

Published by IICTP, an independent technical credential body, on 27 November 2026. The current version of this post is at https://blogs.iictp.org/xbrl-calculation-errors-llm-benchmark-dqc-rules. More guides: blogs.iictp.org.