

VERITROOPER Scout Report — SEC 10K multi — VERITROOPER v10



Model under test: gpt-5.6-sol

Audit executed: 2026-08-26 15:27:11 – 2026-08-26 17:29:18

Report generated: 2026-08-26 17:29:54 (engine v10)

Source: 20260826_152710

Contents (click a section to jump)

Headline Accuracy (bad-test cases excluded from denominator)

Release disposition

Human sign-off

Paired outcome (same questions, both arms)

Per-Category Accuracy & Failure Recovery

Adjudication reconciliation

Test-set coverage

Estimated audit resource usage

Without VERITROOPER — Baseline LLM Performance

With VERITROOPER — Pipeline Performance

Independent Verification

Bottom Line

Deterministic Guardrails

Complete Failure Listing — With VERITROOPER (Pipeline)

Complete Failure Listing — Without VERITROOPER (Baseline)

Audit scope — point-in-time snapshot. This audit is a single point-in-time measurement of **gpt-5.6-sol** answering **889 purpose-generated questions** drawn from **SEC 10K multi — VERITROOPER v10** (SEC_10K_multi.txt), under the recorded retrieval, prompt, and runtime configuration (temperature 0.1). The baseline arm is a standardized BM25 vanilla-RAG reference retrieval.

It measures **only this endpoint × this data set × this configuration at this point in time**. It does **not** measure other deployments, later model or prompt versions, a different data set, or live production traffic — re-audit when any of those change. It is not a general rating of the underlying model.

Headline Accuracy (bad-test cases excluded from denominator)

Scoring set: **889 questions** (of 889 generated; 0 identified as bad tests — questions where the ground truth itself was malformed — excluded symmetrically from both arms' denominators per audit-grade methodology.).

Metric	Without VERITROOPER (vanilla-RAG baseline)	With VERITROOPER (pipeline)
Final verified accuracy	72.89% (648 / 889)	88.08% (783 / 889)
Answers scored incorrect	241	106
Δ vs baseline		+15.19pp
Failure-Recovery Rate (pipeline recovers what % of baseline failures)		64.7%
95% confidence interval (Wilson)	69.9–75.7%	85.8–90.0%

The 95% confidence interval reflects sampling uncertainty on 889 scored items: a point estimate of 100% does not prove a 0% error rate in the field — it means no error was observed in this sample; a tighter bound needs more questions.

Release disposition

Acceptance profile: General enterprise · selected by run configuration (default profile). The audited configuration is measured against this profile's thresholds (below). This is a disposition on the audit **evidence** ; final release additionally requires the responsible person's recorded sign-off and remains the provider's decision.

Configuration	Disposition
Without VERITROOPER (baseline)	FAIL
With VERITROOPER (pipeline)	FAIL

Acceptance criteria & results:

Criterion	Threshold	Baseline	Pipeline
Final verified accuracy	$\geq 95\%$	72.89% FAIL	88.08% FAIL
Answers scored incorrect	$= 0$	241 FAIL	106 FAIL
Unsupported-question accuracy	$\geq 95\%$	60.00% FAIL	87.00% FAIL
Scored sample size	≥ 30	889 PASS	889 PASS
Connection errors (execution failures)	≤ 2	0 PASS	0 PASS

PASS = all criteria met · CONDITIONAL PASS = accuracy target met but a secondary criterion (confirmed errors, unsupported-question handling, or connection errors) missed · FAIL = accuracy target not met · INSUFFICIENT EVIDENCE = sample too small for a reliable call. Connection errors are execution failures (no answer returned), not model verdicts; under this profile a run exceeding the tolerance should re-run the affected item(s) before release. Other selectable profiles (High assurance, Regulated, Safety-critical) raise these bars; all are configurable in the run configuration.

Human sign-off

No human sign-off is on record for these exact results yet. The responsible person records it — name, title, decision, and date (system clock) — at the end of the run in the results screen, or via vt.py signoff; it then binds to this result set's fingerprint.

Paired outcome (same questions, both arms)

Both arms answered the identical scored questions, so the improvement reads pair-by-pair rather than as two separate accuracy figures. **Recovered** = baseline wrong and pipeline right; **regressed** = the reverse.

Paired outcome	Questions
Both correct	627
Recovered (baseline wrong → pipeline correct)	156
Regressed (baseline correct → pipeline wrong)	21
Both wrong	85
Total scored	889

Of the 177 questions where the arms disagreed, 156 favoured the pipeline and 21 favoured the baseline. Exact McNemar two-sided $p = 0.0000$.

Per-Category Accuracy & Failure Recovery

Per-category accuracy with and without VERITROOPER. **Failure-Recovery Rate** measures the fraction of vanilla-RAG baseline failures the pipeline recovers — the most direct signal of where the architecture adds value on this material. Every category is shown, including any where the pipeline matches or underperforms baseline.

Question Type	Questions	Baseline Accuracy	Pipeline Accuracy	Improvement	Failure-Recovery Rate
Negative (hallucination traps)	200	60.00%	87.00%	+27.00pp	76.2%
Precision	235	65.11%	87.66%	+22.55pp	72.0%
Cross-Reference	181	61.88%	76.80%	+14.92pp	46.4%
Conditional	64	92.19%	96.88%	+4.69pp	60.0%
Cause & Effect	112	98.21%	99.11%	+0.89pp	50.0%
Exception	82	100.00%	98.78%	-1.22pp	—
Calculation	15	80.00%	66.67%	-13.33pp	0.0%
All Categories	889	72.89%	88.08%	+15.19pp	64.7%

Adjudication reconciliation

What each arm's automated sweep found, and what review then made of it. "Cleared by verification" are answers the validator flagged but the diagnostic reviewer judged correct — this is reported so you can see where judgement disagreed with the code, and those answers **still count against the model** in the final accuracy, because the deterministic check decides the score. Bad tests (malformed questions) and connection errors (no answer returned — an execution fault) are excluded symmetrically from both arms.

Stage	Baseline	Pipeline
Passed automated review	603	750
Flagged for review	286	139
— cleared by verification (false flags)	45	33
— confirmed model errors	241	106
Final correct / scored	648 / 889	783 / 889
Bad tests excluded (symmetric)	0	0
Connection errors excluded (no answer returned)	0	0
Total generated	889	889

Test-set coverage

What this question panel actually exercised. A point-in-time audit speaks only to the material it covered, so coverage is stated explicitly. "Source passages" are the indexed evidence chunks the questions were generated from — not formal document headings.

Coverage measure	Value
Distinct source passages probed	379
Questions per passage (mean / median)	2.3 / 2
Most / least on a single passage	10 / 1
Passages with only one question	51
Share of questions from the top 5 passages	5%
Adversarial (hallucination-trap) questions	200 of 889 (22%)

- **Category coverage:** Precision (235), Negative (hallucination traps) (200), Cross-Reference (181), Cause & Effect (112), Exception (82), Conditional (64), Calculation (15).

Passage-level counts above are exact. Whole-data-set coverage (% of ALL source passages represented) is not computed here — the data set's total passage count is not recorded in this run; it requires the source manifest (a planned addition).

Estimated audit resource usage

What running this audit consumed. **Measured** values are exact; **estimated** values (tokens, cost) are approximations — the audit is not per-call token-metered — for planning rather than billing.

Measured

- Wall-clock duration: 76 min 3 s.
- Answer calls: 889 baseline + 889 pipeline = 1778.
- Compute locality: MUT **gpt-5.6-sol** (hosted API) · Diagnostic reviewer **gpt-5.6-sol** (hosted API) · Verifier **gemini** (hosted API).

Estimated

- Adjudication calls: ~544 diagnostic reviewer + ~423 independent-verifier review(s) (inferred from the result records, not directly metered).
- Answer-phase tokens: ~2,638,218 (chars÷4 over system prompt + retrieved evidence + question + answer).
- Answer-phase API cost: ~\$26.38 (~\$0.03 per scored question, answer phase only — excludes diagnostic-reviewer / verifier adjudication; at gpt-5.6-sol list price ≈ \$10/M tokens, input+output blended).

Estimated cost is a lower bound — retrieval at answer time may pull more context than the gold evidence chunk, and diagnostic-reviewer / verifier token usage is not included. Local / self-hosted models incur compute, not API cost. Per-call token metering is a planned addition for exact figures.

Without VERITROOPER — Baseline LLM Performance

What the LLM Did On Its Own

With vanilla RAG, the tested LLM answered most questions correctly, but a meaningful minority of answers contained confirmed errors. The failures included unsupported answers to questions that should have been treated as unanswerable, incorrect extraction of values from tables, conflicts with the required numeric answer, and responses that added information beyond the requested result. Malformed questions were excluded from the assessment rather than treated as model mistakes.

Failure Patterns (Baseline)

- **Answered an unanswerable question:** The model supplied a calculated answer even though the available evidence did not support the requested conclusion.
- **Table misread:** The model selected the wrong value from a financial table, indicating difficulty aligning the requested period, row, or column with the relevant figure.
- **Numeric answer conflict:** The model returned a value that did not match the required value, including cases involving totals and financial measures.
- **Computed versus quoted value confusion:** The model mixed calculated, quoted, pre-tax, after-tax, or signed values rather than returning the specific measure requested.
- **Scope and relevance errors:** Some responses included unsupported interpretation or unnecessary contextual material instead of limiting the answer to the requested fact.
- **Evidence-use failure:** In some cases, the response said the evidence did not specify an answer even though the required information was available in the source material.

Baseline Remediation Recommendations

- Strengthen the system prompt so the model must distinguish clearly between values quoted directly from evidence and values it calculates.
 - Require the model to return “UNANSWERABLE” when the requested answer cannot be established from the retrieved material, rather than inferring or reconstructing a result.
 - Require explicit row, column, period, unit, and label alignment before the model answers questions based on financial tables.
 - Add an application-level numeric consistency check that compares the proposed answer with the cited source value and verifies units, signs, tax treatment, and reporting period.
 - Require concise answers that address only the requested fact unless explanation is explicitly requested.
 - Determine whether each evidence-use failure resulted from a retrieval miss or from the model not using retrieved evidence. Adjust retrieval only where the relevant passage was not supplied.
 - Route unsupported calculations, ambiguous table lookups, and material financial totals to human review before release.
 - Use targeted prompt examples and regression tests covering unanswerable questions, table extraction, period selection, and computed-versus-quoted values.
-

With VERITROOPER — Pipeline Performance

What the LLM Did With VERITROOPER

When wrapped by VERITROOPER, the tested LLM had a small minority of confirmed residual errors, and the broader baseline patterns were not present as confirmed model failures in this audit panel. The remaining confirmed issue was a numeric conflict involving the total value of asset-backed securities. Other residual cases were malformed or excluded tests and therefore were not treated as model errors.

Pipeline Remediation Recommendations

- Configure the model to show the source components used for financial totals before producing the final answer.
 - Add an application-level reconciliation step that checks whether an aggregate matches the relevant table entries, reporting period, and units.
 - Require a citation to the exact table row or disclosure supporting any material total.
 - Route unresolved aggregate-value conflicts to human review and sign-off.
 - Maintain a focused regression test for asset-backed securities totals and similar table-based aggregation questions.
-

Independent Verification

An independent verifier audited the diagnostic decisions for both the pipeline and baseline arms under the same standard. This symmetric review supports an honest comparison by applying equivalent scrutiny to answers from both configurations.

Bottom Line

For the tested panel, the VERITROOPER-wrapped configuration is the stronger deployment choice because it left fewer answers requiring post-generation correction and the baseline failure patterns were largely absent from the pipeline arm. Human review should remain in place for material financial totals and unresolved table-based calculations, particularly where an incorrect aggregate could affect a business decision.

Deterministic Guardrails

Code-level checks you can deploy in your own stack to catch the failure patterns found in this run — no model retraining required. Because they are deterministic, they behave identically on every run. **Block** = the check is authoritative and can reject or correct the answer automatically; **Flag** = the check detects the error class and routes it to human review or abstention. This complements the remediation recommendations above — it does not replace them.

Block — reject or correct automatically:

- **Unit/format normalization** — Parse the value AND its unit, convert to a canonical unit/scale (% , \$, thousands vs millions) before comparison, and reject unit or scale mismatches. (Addresses: Misinterpreted, Out of scope, Table misread, Unit/format mismatch, Wrong calc/quote, Wrong period · 74 cases this run.)
- **Grounding check** — Require every figure and factual claim in the answer to appear in — or be directly entailed by — the retrieved evidence; reject answers with unsupported spans (the model leaning on its own knowledge over the source). (Addresses: Meta leak, Misinterpreted, Out of scope, Table misread, Unit/format mismatch, Wrong period · 18 cases this run.)
- **Recompute check** — Recompute the arithmetic from the source figures and reject any answer whose stated result differs beyond a rounding tolerance. (Addresses: Math error, Misinterpreted, Out of scope, Unit/format mismatch, Wrong calc/quote · 18 cases this run.)
- **Numeric tolerance** — Normalize both values to a common precision and accept only within an explicit rounding tolerance (e.g. ± 0.5 of the least-significant digit). (Addresses: Misinterpreted, Out of scope, Table misread, Unit/format mismatch · 8 cases this run.)
- **Quote-vs-compute guard** — Decide whether the answer should be a verbatim quote or a computed value; validate a quote as a substring of the source and a computed value by recomputation. (Addresses: Wrong calc/quote · 2 cases this run.)

Flag — detect and route to review:

- **Cell-provenance check** — Confirm the answered value comes from the exact row + column the question names (row/column-header match); flag row/column drift. (Addresses: Out of scope, Table misread, Wrong entity · 5 cases this run.)
- **Direction/polarity check** — Compare the answer's polarity and direction (yes/no, increase/decrease, permitted/prohibited) against the source statement; flag reversals. (Addresses: deterministic:direction conflict · 4 cases this run.)
- **Evidence-coverage check** — Before answering, verify the retrieved context actually contains a span that supports the question; when coverage is below a set threshold, route to broader / multi-query retrieval or abstain rather than answering from a thin context. (Addresses: Table misread · 2 cases this run.)
- **Period-match check** — Extract the period/date from the question and require the answer's figure to come from that period; flag period mismatches. (Addresses: Out of scope · 1 case this run.)
- **Entity-match check** — Require the answer to reference the same entity named in the question (exact or alias match); flag cross-entity substitutions. (Addresses: Out of scope · 1 case this run.)

- **Abstention guard** — When evidence coverage is below a set threshold, require the model to abstain or hedge; flag confident answers made over weak or absent evidence. (Addresses: Certainty mismatch · 1 case this run.)

Case counts are per flagged answer and can overlap — a single answer may exhibit more than one failure pattern — so they need not sum to the number of failed questions. Patterns without a listed guard are semantic reasoning slips with no clean deterministic check; the remediation recommendations above address those.

Complete Failure Listing — With VERITROOPER (Pipeline)

71 confirmed model errors + 35 excluded bad tests (malformed question, not counted against the model). Each entry shows the full question, the model's full answer, the ground truth, the diagnostic reviewer's diagnosis, and (where available) the Verifier's independent audit verdict.

Failure shape at a glance — With VERITROOPER (Pipeline)

Category	Scored incorrect	Model error	Disputed (adjudicators cleared)	Bad test	Other	Review recommended
Cross-Reference	42	8	19	15	0	3
Precision	29	9	13	7	0	3
Negative (hallucination traps)	26	17	0	9	0	2
Calculation	5	1	0	4	0	0
Conditional	2	2	0	0	0	0
Exception	1	1	0	0	0	0
Cause & Effect	1	0	1	0	0	0
TOTAL	106	38	33	35	0	8

Scored incorrect is the count that reaches the headline — the deterministic check scored these answers wrong. The next four columns say what adjudication then found, and they sum to it. **Disputed** means the diagnostic reviewer (no independent Verifier ran on this audit) judged the answer correct; those answers still count against the model in the headline, because the deterministic check decides the score and the model verdicts are advisory.

Severity mix (confirmed errors): 69 HIGH, 2 MEDIUM. Severity is a domain-agnostic heuristic from the error type — a confidently-wrong answer outranks a missing one; re-weight for your risk context.

Failure 1 — `SEC\_10K\_multi\_neg\_0516\_01` · LOW confidence

BAD QUESTION

- **Category:** Negative (hallucination traps)
- **Question:** What is the total amount of deposits that are covered by deposit insurance protection at JPMorgan Chase as of December 31, 2025?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** Deposits covered by insurance protection totaled approximately **\$1,000.7 billion** as of December 31, 2025 (\$2,559.3 billion total deposits less \$1,558.6 billion uninsured).
- **Diagnostic reviewer diagnosis:** **Negative-answer specialist** bad test
- **Diagnostic reviewer root cause:** evidence contains the answer; UNANSWERABLE label was wrong
- **Verifier verdict:** bad test (confidence: low)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer; Unsupported number: '\$1,000.7' does not appear in the evidence
- **Review recommended:** yes (Verifier flagged grey-area)

Failure 2 — `SEC\_10K\_multi\_neg\_0714\_02` · LOW confidence

BAD QUESTION

- **Category:** Negative (hallucination traps)
- **Question:** What is the name of the specific accounting standard that the new guidance is intended to clarify?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** The standard is **Revenue from Contracts with Customers**.
- **Diagnostic reviewer diagnosis:** **Negative-answer specialist** bad test
- **Diagnostic reviewer root cause:** evidence contains the answer; UNANSWERABLE label was wrong
- **Verifier verdict:** bad test (confidence: low)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer
- **Review recommended:** yes (Verifier flagged grey-area)

Failure 3 — `SEC\_10K\_multi\_0064\_00` · LOW confidence

BAD QUESTION

- **Category:** Cross-Reference
- **Question:** What is the total amount of cash and cash equivalents at the beginning of the period for the year ended January 25, 2026, and the year ended January 26, 2025?
- ✓ **Ground truth:** 10,605 and 8,589
- ✗ **Model answered:** Cash and cash equivalents at the beginning of the period were \$8,589 million for 2026 and \$7,280 million for 2025, totaling \$15,869 million.
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** bad test
- **Diagnostic reviewer root cause:** The model correctly read the cash-flow row for beginning-of-period cash (\$8,589 million and \$7,280 million) and summed the two values to \$15,869 million; the validator incorrectly treated the derived total as unsupported because it was not stated verbatim.
- **Verifier verdict:** bad test (confidence: low)
- **Validator flags:** Unsupported number: '\$15,869' does not appear in the evidence
- **Review recommended:** yes (Verifier flagged grey-area)

Failure 4 — `SEC\_10K\_multi\_1588\_00` · LOW confidence

UNIT/FORMAT MISMATCH

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Cross-Reference
- **Question:** What is the total amount of assets held for sale as of December 31, 2025, and how does it compare to the total amount of liabilities held for sale?
- ✓ **Ground truth:** The total amount of assets held for sale as of December 31, 2025, is \$5,342 million, and the total amount of liabilities held for sale is \$2,570 million.
- ✗ **Model answered:** As of December 31, 2025, assets held for sale totaled \$5.342 billion, compared with liabilities held for sale of \$2.570 billion; assets exceeded liabilities by \$2.772 billion.
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** false positive
- **Diagnostic reviewer root cause:** Validator could not recognize the unit conversion from \$5,342 million to \$5.342 billion (and likewise \$2,570 million to \$2.570 billion); the model also correctly computed the \$2.772 billion difference.
- **Verifier verdict:** false positive (confidence: low)
- **Validator flags:** Unsupported number: '\$5.342' does not appear in the evidence
- **Review recommended:** yes (Verifier flagged grey-area)

Failure 5 — `SEC\_10K\_multi\_0216\_01` · LOW confidence

WRONG CALC/QUOTE

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Cross-Reference
- **Question:** What is the cumulative total return for JPMorgan Chase & Co. common stock over the five-year period ending December 31, 2025, and how does it compare to the S&P 500 Index?
- ✓ **Ground truth:** The cumulative total return for JPMorgan Chase & Co. common stock was \$289.18, which is higher than the S&P 500 Index return of \$196.12.
- ✗ **Model answered:** JPMorganChase common stock returned 189.18% cumulatively, versus 96.12% for the S&P 500 Index, outperforming it by 93.06 percentage points.
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** false positive
- **Diagnostic reviewer root cause:** The model converted the table's indexed ending values, based on an initial \$100 investment, into percentage gains: \$289.18 corresponds to a 189.18% return and \$196.12 to a 96.12% return. The validator only matched the quoted dollar/index values and did not recognize this equivalent calculation.
- **Verifier verdict:** false positive (confidence: low)
- **Validator flags:** Unsupported number: '189.18%' does not appear in the evidence
- **Review recommended:** yes (Verifier flagged grey-area)

Failure 6 — `SEC\_10K\_multi\_1173\_01` · HIGH confidence

WRONG CALC/QUOTE

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Cross-Reference
- **Question:** What is the weighted-average extension of loans for single modifications, and how does it compare to the extension for interest rate reduction and term extension?
- ✓ **Ground truth:** 20 months for single modifications, and 36 months for interest rate reduction and term extension.

- **Model answered:** Single-modification term extensions averaged 20 months, compared with 36 months for interest rate reduction and term extension-16 months longer.
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** false positive
- **Diagnostic reviewer root cause:** The model correctly reported both extensions and additionally calculated their difference ($36 - 20 = 16$ months). The validator only checked for numbers explicitly present in the evidence and failed to recognize this valid derived comparison.
- **Verifier verdict:** false positive (confidence: high)
- **Validator flags:** Unsupported number: '16' does not appear in the evidence

Failure 7 — `SEC\_10K\_multi\_1501\_01` · LOW confidence

UNIT/FORMAT MISMATCH

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Precision
- **Question:** What is the decrease in deposits with banks due to change in rate?
- **Ground truth:** (2,792)
- **Model answered:** The decrease in deposits with banks due to the change in rate was **\$2.792 billion**.
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** false positive
- **Diagnostic reviewer root cause:** The model correctly converted the table value of \$(2,792) million into a \$2.792 billion decrease, but the validator failed to recognize the unit conversion and parenthetical negative/decrease notation.
- **Verifier verdict:** false positive (confidence: low)
- **Validator flags:** Unsupported number: '\$2.792' does not appear in the evidence
- **Review recommended:** yes (Verifier flagged grey-area)

Failure 8 — `SEC\_10K\_multi\_0747\_01` · — confidence

DETERMINISTIC:GT NUMBER CONFLICT

- **Severity:** HIGH — an incorrect quantitative result
- **Category:** Precision
- **Question:** What is the total value of asset-backed securities?
- **Ground truth:** 11,316
- **Model answered:** The total value of asset-backed securities is **2,823**.
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** model error
- **Diagnostic reviewer root cause:** [code:gt_number_conflict] Short numeric GT; response asserts a different quantity
- **Verifier verdict:** — (confidence: —)
- **Validator flags:** (none)

Failure 9 — `SEC\_10K\_multi\_0072\_00` · LOW confidence

VALIDATOR OVER-FIRED

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Precision
- **Question:** What percentage of accounts receivable balance as of January 25, 2026, was accounted for by three direct customers?
- **Ground truth:** 25
- **Model answered:** The three direct customers accounted for a combined 56% of accounts receivable as of January 25, 2026.

- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** false positive
- **Diagnostic reviewer root cause:** pre-Doctor equivalence: numeric match: model has 25 ~ GT 25
- **Verifier verdict:** false positive (confidence: low)
- **Validator flags:** Unsupported number: '56%' does not appear in the evidence
- **Review recommended:** yes (Verifier flagged grey-area)

Failure 10 — `SEC\_10K\_multi\_1159\_00` · HIGH confidence

BAD QUESTION

- **Category:** Precision
- **Question:** What is the total retained Secured by real estate loans for 2025?
- ✓ **Ground truth:** 105,130
- ✗ **Model answered:** The total retained secured by real estate loans for 2025 was **\$165.545 billion**.
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** bad test
- **Diagnostic reviewer root cause:** The model correctly read the 2025 Total column value of \$165,545 million and converted it to \$165.545 billion; the validator failed to recognize the unit conversion, while the ground truth appears to have read the 2025 Multifamily value instead.
- **Verifier verdict:** bad test (confidence: high)
- **Validator flags:** Unsupported number: '\$165.545' does not appear in the evidence

Failure 11 — `SEC\_10K\_multi\_1600\_00` · LOW confidence

UNIT/FORMAT MISMATCH

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Precision
- **Question:** What is the amount of loss carryforwards available to reduce future taxable income as of December 31, 2025?
- ✓ **Ground truth:** 1,600 million
- ✗ **Model answered:** As of December 31, 2025, the Company had **\$1.6 billion** in loss carryforwards available to reduce future taxable income.
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** false positive
- **Diagnostic reviewer root cause:** Validator failed to recognize unit conversion: \$1.6 billion is numerically equivalent to the evidence's \$1,600 million.
- **Verifier verdict:** false positive (confidence: low)
- **Validator flags:** Unsupported number: '\$1.6' does not appear in the evidence
- **Review recommended:** yes (Verifier flagged grey-area)

Failure 12 — `SEC\_10K\_multi\_neg\_0034\_04` · HIGH confidence

BAD QUESTION

- **Category:** Negative (hallucination traps)
- **Question:** What is the total number of notes issued by Apple Inc. in 2022?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** Apple Inc. issued four notes in 2022.
- **Diagnostic reviewer diagnosis:** **Negative-answer specialist** bad test
- **Diagnostic reviewer root cause:** evidence contains the answer; UNANSWERABLE label was wrong
- **Verifier verdict:** bad test (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 13 — `SEC\_10K\_multi\_neg\_1265\_02` · — confidence

DETERMINISTIC:ANSWERED UNANSWERABLE

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What is the average annual increase in goodwill for the Commercial & Investment Bank segment?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** The average annual increase in goodwill was **\$4 million** from 2023 to 2025.
- **Diagnostic reviewer diagnosis:** **Negative-answer specialist** model error
- **Diagnostic reviewer root cause:** [code:answered_unanswerable] GT is UNANSWERABLE but the response asserts a concrete answer
- **Verifier verdict:** — (confidence: —)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer; Unsupported number: '\$4' does not appear in the evidence

Failure 14 — `SEC\_10K\_multi\_0060\_01` · HIGH confidence

UNIT/FORMAT MISMATCH

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Precision
- **Question:** What is the amount of cash used in financing activities for fiscal year 2026?
- ✓ **Ground truth:** 48,474
- ✗ **Model answered:** Net cash used in financing activities for fiscal year 2026 was \$48.474 billion.
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** false positive
- **Diagnostic reviewer root cause:** Validator failed to account for unit conversion: the table reports \$48,474 million, which is exactly \$48.474 billion.
- **Verifier verdict:** false positive (confidence: high)
- **Validator flags:** Unsupported number: '\$48.474' does not appear in the evidence

Failure 15 — `SEC\_10K\_multi\_neg\_1274\_01` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What is the average contractual service fee included in income?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** The average contractual service fee included in income was **1,610.33**.
- **Diagnostic reviewer diagnosis:** **Negative-answer specialist** model error
- **Diagnostic reviewer root cause:** model hallucinated answer not in evidence
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer; Unsupported number: '1,610.33' does not appear in the evidence

Failure 16 — `SEC\_10K\_multi\_1403\_01` · — confidence

DETERMINISTIC:GT NUMBER CONFLICT

- **Severity:** HIGH — an incorrect quantitative result
- **Category:** Cross-Reference
- **Question:** How do the unsettled repurchase and securities loaned agreements compare to the unsettled resale and securities borrowed agreements?
- ✓ **Ground truth:** 52,308 is less than 136,841.
- ✗ **Model answered:** Unsettled repurchase and securities loaned agreements were lower than unsettled resale and securities borrowed agreements: 52,895 versus 137,072 in Col. 5, and 66,986 versus 115,939 in Col. 6.
- **Diagnostic reviewer diagnosis:** General diagnostic specialist model error
- **Diagnostic reviewer root cause:** [code:gt_number_conflict] Short numeric GT; response asserts a different quantity
- **Verifier verdict:** — (confidence: —)
- **Validator flags:** (none)

Failure 17 — `SEC\_10K\_multi\_0072\_01` · HIGH confidence

UNIT/FORMAT MISMATCH

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Precision
- **Question:** What was the total inventories as of January 26, 2025?
- ✓ **Ground truth:** 10,080
- ✗ **Model answered:** Total inventories as of January 26, 2025 were **\$10.08 billion**.
- **Diagnostic reviewer diagnosis:** General diagnostic specialist false positive
- **Diagnostic reviewer root cause:** Validator could not recognize the unit conversion from the table's \$10,080 million to the model's equivalent \$10.08 billion.
- **Verifier verdict:** false positive (confidence: high)
- **Validator flags:** Unsupported number: '\$10.08' does not appear in the evidence

Failure 18 — `SEC\_10K\_multi\_0318\_01` · HIGH confidence

BAD QUESTION

- **Category:** Cross-Reference
- **Question:** How does the bank's revenue from card income compare to its total net revenue in 2025?
- ✓ **Ground truth:** The bank's revenue from card income was \$2,436 million, which is part of the total net revenue of \$53,766 million in 2025.
- ✗ **Model answered:** In 2025, card income was \$2.436 billion, compared with total net revenue of \$78.454 billion, or approximately 3.1% of the total.
- **Diagnostic reviewer diagnosis:** General diagnostic specialist bad test
- **Diagnostic reviewer root cause:** The model converted \$2,436 million to \$2.436 billion and correctly derived total net revenue as \$78.454 billion by adding noninterest revenue of \$53.766 billion and net interest income of \$24.688 billion; the validator also failed to recognize the unit conversion.
- **Verifier verdict:** bad test (confidence: high)
- **Validator flags:** Unsupported number: '\$2.436' does not appear in the evidence

Failure 19 — `SEC\_10K\_multi\_1125\_00` · HIGH confidence

UNIT/FORMAT MISMATCH

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Precision
- **Question:** What is the total retained loans for 2024?
- ✓ **Ground truth:** 376,334
- ✗ **Model answered:** The total retained loans for 2024 were **\$376.334 billion**.
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** false positive
- **Diagnostic reviewer root cause:** The model converted the table's \$376,334 million figure to \$376.334 billion, which is numerically equivalent, but the validator relied on literal number matching and did not account for the unit conversion.
- **Verifier verdict:** false positive (confidence: high)
- **Validator flags:** Unsupported number: '\$376.334' does not appear in the evidence

Failure 20 — `SEC\_10K\_multi\_0080\_00` · HIGH confidence

UNIT/FORMAT MISMATCH

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Cross-Reference
- **Question:** What is the total amount of current income taxes for the year ended January 25, 2026, and how does it compare to the year ended January 26, 2025?
- ✓ **Ground truth:** The total current income taxes for the year ended January 25, 2026, is \$22,807, which is \$7,184 more than the \$15,623 for the year ended January 26, 2025.
- ✗ **Model answered:** Total current income taxes were \$22.807 billion for 2026, compared with \$15.623 billion for 2025—an increase of \$7.184 billion, or approximately 46%.
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** false positive
- **Diagnostic reviewer root cause:** The model correctly converted the table's figures from millions to billions: \$22,807 million equals \$22.807 billion, and \$15,623 million equals \$15.623 billion. The validator failed to recognize this unit conversion because the exact string '\$22.807' does not appear in the evidence.
- **Verifier verdict:** false positive (confidence: high)
- **Validator flags:** Unsupported number: '\$22.807' does not appear in the evidence

Failure 21 — `SEC\_10K\_multi\_neg\_0767\_00` · — confidence

DETERMINISTIC:ANSWERED UNANSWERABLE

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What is the average price of oil in BBL?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** The average oil commodity forward price was **\$202 per BBL**.
- **Diagnostic reviewer diagnosis:** **Negative-answer specialist** model error
- **Diagnostic reviewer root cause:** [code:answered_unanswerable] GT is UNANSWERABLE but the response asserts a concrete answer
- **Verifier verdict:** — (confidence: —)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 22 — `SEC\_10K\_multi\_1155\_01` · HIGH confidence

AMBIGUOUS QUESTION

- **Category:** Cross-Reference
- **Question:** What is the total noninvestment-grade loans for 2025 and 2024, and what percentage of total retained loans are criticized nonaccrual loans for each year?
- ✓ **Ground truth:** The total noninvestment-grade loans for 2025 is \$46,670 and for 2024 is \$48,152. The percentage of total retained loans that are criticized nonaccrual loans for 2025 is 1.01% and for 2024 is 0.89%.
- ✗ **Model answered:** Total noninvestment-grade loans were \$251,003 in 2025 and \$218,726 in 2024. Criticized nonaccrual loans represented 0.6% of total retained loans in both years.
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** bad test
- **Diagnostic reviewer root cause:** The model selected the rightmost aggregate column pair (251,003/218,726 and 0.56%/0.57%, rounded to 0.6%) instead of the first loan-category column pair intended by the ground truth (46,670/48,152 and 1.01%/0.89%).
- **Verifier verdict:** bad test (confidence: high)
- **Validator flags:** Unsupported number: '0.6%' does not appear in the evidence

Failure 23 — `SEC\_10K\_multi\_neg\_1595\_03` · — confidence

DETERMINISTIC:ANSWERED UNANSWERABLE

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What is the total value of the Company's other postretirement benefit plans' assets as of December 31, 2025?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** The fair value of the Company's other postretirement benefit plan assets was **\$67 million** as of December 31, 2025.
- **Diagnostic reviewer diagnosis:** **Negative-answer specialist** model error
- **Diagnostic reviewer root cause:** [code:answered_unanswerable] GT is UNANSWERABLE but the response asserts a concrete answer
- **Verifier verdict:** — (confidence: —)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 24 — `SEC\_10K\_multi\_0368\_00` · HIGH confidence

WRONG CALC/QUOTE

- **Category:** Cross-Reference
- **Question:** What were the total assets under custody in Fixed Income and Equity Markets as of December 31, 2025?
- ✓ **Ground truth:** 18,322 billion and 17,954 billion
- ✗ **Model answered:** Total assets under custody were **\$36,276 billion** as of December 31, 2025 (\$18,322 billion Fixed Income plus \$17,954 billion Equity).
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** bad test
- **Diagnostic reviewer root cause:** The model correctly reported both category figures and additionally computed their sum; the validator rejected the derived \$36,276 billion because it does not appear verbatim in the evidence.
- **Verifier verdict:** bad test (confidence: high)
- **Validator flags:** Unsupported number: '\$36,276' does not appear in the evidence

Failure 25 — `SEC\_10K\_multi\_0551\_00` · HIGH confidence

UNIT/FORMAT MISMATCH

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Precision
- **Question:** What was the total nonaccrual loans at December 31, 2025?
- ✓ **Ground truth:** 9,857
- ✗ **Model answered:** Total nonaccrual loans at December 31, 2025 were **\$9.857 billion**.
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** false positive
- **Diagnostic reviewer root cause:** Validator failed to recognize a unit conversion: the table reports \$9,857 million, which is exactly \$9.857 billion.
- **Verifier verdict:** false positive (confidence: high)
- **Validator flags:** Unsupported number: '\$9.857' does not appear in the evidence

Failure 26 — `SEC\_10K\_multi\_neg\_0034\_08` · — confidence

DETERMINISTIC:ANSWERED UNANSWERABLE

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What is the average interest rate of the notes issued by Apple Inc. in 2022?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** The average interest rate of Apple's notes issued in 2022 was **3.6625%**.
- **Diagnostic reviewer diagnosis:** **Negative-answer specialist** model error
- **Diagnostic reviewer root cause:** [code:answered_unanswerable] GT is UNANSWERABLE but the response asserts a concrete answer
- **Verifier verdict:** — (confidence: —)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer; Unsupported number: '3.6625%' does not appear in the evidence

Failure 27 — `SEC\_10K\_multi\_0747\_00` · — confidence

DETERMINISTIC:GT NUMBER CONFLICT

- **Severity:** HIGH — an incorrect quantitative result
- **Category:** Precision
- **Question:** What is the total value of equity securities?
- ✓ **Ground truth:** 417,428
- ✗ **Model answered:** The total value of equity securities is **131,728**.
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** model error
- **Diagnostic reviewer root cause:** [code:gt_number_conflict] Short numeric GT; response asserts a different quantity
- **Verifier verdict:** — (confidence: —)
- **Validator flags:** (none)

Failure 28 — `SEC\_10K\_multi\_1354\_00` · HIGH confidence

BAD QUESTION

- **Category:** Cross-Reference

- **Question:** What was the total income tax expense for the year ended December 31, 2025, and what portion of it was due to tax benefits?
- ✓ **Ground truth:** 15,547, 9,936
- ✗ **Model answered:** Total income tax expense was **\$15.547 billion** in 2025, including **\$629 million in tax benefits**, or approximately **4.0%** of the total.
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** bad test
- **Diagnostic reviewer root cause:** The model correctly converted \$15,547 million to \$15.547 billion and used the explicitly stated \$629 million tax-benefit figure; the validator failed to recognize the unit conversion, while the reference answer incorrectly selected \$9,936 million from the 'Total current income tax expense' row.
- **Verifier verdict:** bad test (confidence: high)
- **Validator flags:** Unsupported number: '\$15.547' does not appear in the evidence

Failure 29 — `SEC\_10K\_multi\_0518\_00` · — confidence

DETERMINISTIC:GT NUMBER CONFLICT

- **Severity:** HIGH — an incorrect quantitative result
- **Category:** Cross-Reference
- **Question:** What is the total interest-bearing deposits in non-U.S. offices in the third year, given the total interest-bearing deposits in the first year and the growth rate?
- ✓ **Ground truth:** 408,419
- ✗ **Model answered:** The total interest-bearing deposits in non-U.S. offices in the third year were **\$482,735**.
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** model error
- **Diagnostic reviewer root cause:** [code:gt_number_conflict] Short numeric GT; response asserts a different quantity
- **Verifier verdict:** — (confidence: —)
- **Validator flags:** (none)

Failure 30 — `SEC\_10K\_multi\_1265\_01` · HIGH confidence

BAD QUESTION

- **Category:** Cross-Reference
- **Question:** What is the net liquidation loss for securitized loans in 2025, and how does it compare to the net liquidation loss in 2024?
- ✓ **Ground truth:** The net liquidation loss for securitized loans in 2025 is \$5,233 million, which is higher than the net liquidation loss in 2024 of \$2,329 million.
- ✗ **Model answered:** The net liquidation loss for securitized loans was \$301 million in 2025, compared with \$89 million in 2024—an increase of \$212 million.
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** bad test
- **Diagnostic reviewer root cause:** The ground truth read the '90 days past due' totals (\$5,233 million and \$2,329 million) as net liquidation losses. The model correctly used the net liquidation loss columns (\$301 million and \$89 million) and calculated the \$212 million increase; the validator incorrectly rejected this derived difference because it was not stated verbatim.
- **Verifier verdict:** bad test (confidence: high)
- **Validator flags:** Unsupported number: '\$212' does not appear in the evidence

Failure 31 — `SEC\_10K\_multi\_0231\_00` · — confidence

DETERMINISTIC:GT NUMBER CONFLICT

- **Severity:** HIGH — an incorrect quantitative result
- **Category:** Precision

- **Question:** What is the total credit provided and capital raised for U.S. small businesses?
- ✓ **Ground truth:** \$2.9 trillion
- ✗ **Model answered:** The total credit provided and capital raised for U.S. small businesses was **\$33 billion** in 2025.
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** model error
- **Diagnostic reviewer root cause:** [code:gt_number_conflict] Short numeric GT; response asserts a different quantity
- **Verifier verdict:** — (confidence: —)
- **Validator flags:** (none)

Failure 32 — `SEC\_10K\_multi\_0574\_01` · HIGH confidence

UNIT/FORMAT MISMATCH

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Precision
- **Question:** What is the total number of government guaranteed loans at December 31, 2025?
- ✓ **Ground truth:** 1,159
- ✗ **Model answered:** The total government guaranteed loans at December 31, 2025, were \$1.159 billion.
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** false positive
- **Diagnostic reviewer root cause:** Validator failed to recognize the unit conversion: the table reports \$1,159 million, which is exactly \$1.159 billion.
- **Verifier verdict:** false positive (confidence: high)
- **Validator flags:** Unsupported number: '\$1.159' does not appear in the evidence

Failure 33 — `SEC\_10K\_multi\_0911\_00` · HIGH confidence

BAD QUESTION

- **Category:** Cross-Reference
- **Question:** What is the total amount for Insurance and what is the total amount for Central Govt, and how do these two totals compare?
- ✓ **Ground truth:** The total amount for Insurance is \$25,031 and the total amount for Central Govt is \$15,164. These two totals are not comparable as they are different categories.
- ✗ **Model answered:** Insurance totaled 25,031, while Central Govt totaled 15,164. Insurance was higher by 9,867.
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** bad test
- **Diagnostic reviewer root cause:** The model reasonably interpreted “compare” as requesting a numerical comparison and subtracted 15,164 from 25,031 to obtain 9,867, whereas the ground truth treats the categories as non-comparable.
- **Verifier verdict:** bad test (confidence: high)
- **Validator flags:** Unsupported number: '9,867' does not appear in the evidence

Failure 34 — `SEC\_10K\_multi\_neg\_0744\_03` · — confidence

DETERMINISTIC:ANSWERED UNANSWERABLE

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What is the total amount of accounts payable and other liabilities measured at fair value on a recurring basis?
- ✓ **Ground truth:** UNANSWERABLE

- **✗ Model answered:** The total accounts payable and other liabilities measured at fair value on a recurring basis was **\$6.660 billion**.
- **Diagnostic reviewer diagnosis:** **Negative-answer specialist** model error
- **Diagnostic reviewer root cause:** [code:answered_unanswerable] GT is UNANSWERABLE but the response asserts a concrete answer
- **Verifier verdict:** — (confidence: —)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer; Unsupported number: '\$6.660' does not appear in the evidence

Failure 35 — `SEC\_10K\_multi\_0082\_01` · HIGH confidence

UNIT/FORMAT MISMATCH

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Precision
- **Question:** What is the amount of cash paid for income taxes (net of refunds) for the fiscal year ended January 25, 2026?
- **✓ Ground truth:** \$20,288
- **✗ Model answered:** The cash paid for income taxes, net of refunds, was **\$20.288 billion** for the fiscal year ended January 25, 2026.
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** false positive
- **Diagnostic reviewer root cause:** Validator failed to recognize the unit conversion: the table reports \$20,288 million, which is exactly \$20.288 billion.
- **Verifier verdict:** false positive (confidence: high)
- **Validator flags:** Unsupported number: '\$20.288' does not appear in the evidence

Failure 36 — `SEC\_10K\_multi\_0320\_00` · HIGH confidence

WRONG CALC/QUOTE

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Cross-Reference
- **Question:** What were the total taxable-equivalent adjustments from income tax credits for the years ended December 31, 2023, 2024, and 2025?
- **✓ Ground truth:** \$4.0 billion, \$2.8 billion, and \$2.9 billion
- **✗ Model answered:** The total taxable-equivalent adjustments for 2023, 2024, and 2025 were **\$9.7 billion** (\$4.0 billion, \$2.8 billion, and \$2.9 billion, respectively).
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** false positive
- **Diagnostic reviewer root cause:** The model correctly reported all three annual adjustments and additionally summed them to \$9.7 billion, interpreting “total” as an aggregate. The validator rejected the accurately computed sum because it does not appear verbatim in the evidence.
- **Verifier verdict:** false positive (confidence: high)
- **Validator flags:** Unsupported number: '\$9.7' does not appear in the evidence

Failure 37 — `SEC\_10K\_multi\_0109\_00` · — confidence

DETERMINISTIC:DIRECTION CONFLICT

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Conditional
- **Question:** If JPMorganChase is a large accelerated filer, does it have to file reports pursuant to Section 13 or 15(d) of the Act?

- ✓ **Ground truth:** No
- ✗ **Model answered:** Yes. JPMorganChase is a large accelerated filer and indicated that it is required to file reports pursuant to Section 13 or 15(d) of the Act.
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** model error
- **Diagnostic reviewer root cause:** [code:direction_conflict] GT is 'no' but response leads 'yes'
- **Verifier verdict:** — (confidence: —)
- **Validator flags:** Direction mismatch: GT negative, response affirmative

Failure 38 — `SEC\_10K\_multi\_0911\_01` · HIGH confidence

BAD QUESTION

- **Category:** Cross-Reference
- **Question:** What is the total amount for Transportation and what is the total amount for Financial Markets Infrastructure, and how do these two totals compare?
- ✓ **Ground truth:** The total amount for Transportation is \$20,861 and the total amount for Financial Markets Infrastructure is \$5,734. These two totals are not comparable as they are different categories.
- ✗ **Model answered:** Transportation totals 20,861, while Financial Markets Infrastructure totals 5,734. Transportation is 15,127 higher, or about 3.6 times as much.
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** bad test
- **Diagnostic reviewer root cause:** The model interpreted “how do these totals compare?” as requesting an arithmetic difference and ratio, while the reference answer treated the categories as non-comparable; the validator also failed to recognize 3.6 as a valid derived ratio.
- **Verifier verdict:** bad test (confidence: high)
- **Validator flags:** Unsupported number: '3.6' does not appear in the evidence

Failure 39 — `SEC\_10K\_multi\_1297\_01` · HIGH confidence

UNIT/FORMAT MISMATCH

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Cross-Reference
- **Question:** What is the net rental expense for the year ended December 31, 2025, and how does it compare to the gross rental expense?
- ✓ **Ground truth:** Net rental expense: \$2,360, Gross rental expense: \$2,388
- ✗ **Model answered:** Net rental expense was \$2.360 billion in 2025, compared with gross rental expense of \$2.388 billion—a \$28 million reduction from sublease rental income.
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** false positive
- **Diagnostic reviewer root cause:** Validator could not recognize the unit conversion: \$2,360 million is equivalent to \$2.360 billion, and \$2,388 million is equivalent to \$2.388 billion.
- **Verifier verdict:** false positive (confidence: high)
- **Validator flags:** Unsupported number: '\$2.360' does not appear in the evidence

Failure 40 — `SEC\_10K\_multi\_0929\_01` · HIGH confidence

BAD QUESTION

- **Category:** Calculation
- **Question:** If the gross long notional derivative contracts are \$500 and the gross short notional derivative contracts are \$300, what is the total?
- ✓ **Ground truth:** 200
- ✗ **Model answered:** 800

- **Diagnostic reviewer diagnosis:** Calculation specialist bad test
- **Diagnostic reviewer root cause:** model matches evidence formula; GT is wrong
- **Verifier verdict:** bad test (confidence: high)
- **Validator flags:** Number mismatch: expected 200, not found in response

Failure 41 — `SEC\_10K\_multi\_1501\_00` · HIGH confidence

BAD QUESTION

- **Category:** Precision
- **Question:** What is the amount of noninterest-bearing liabilities funding interest-earning assets?
- ✓ **Ground truth:** 284,913
- ✗ **Model answered:** Noninterest-bearing liabilities funding interest-earning assets were \$701.189 billion in 2024 and \$731.939 billion in 2023.
- **Diagnostic reviewer diagnosis:** General diagnostic specialist bad test
- **Diagnostic reviewer root cause:** The model correctly identified the table row derived as total interest-earning assets minus total interest-bearing liabilities: \$701,189 million in 2024 and \$731,939 million in 2023. The validator failed to normalize the equivalent billion-denominated formatting, while the ground truth selected an unrelated first-row value.
- **Verifier verdict:** bad test (confidence: high)
- **Validator flags:** Unsupported number: '\$701.189' does not appear in the evidence

Failure 42 — `SEC\_10K\_multi\_neg\_0034\_00` · HIGH confidence

BAD QUESTION

- **Category:** Negative (hallucination traps)
- **Question:** What is the total number of notes issued by Apple Inc. in 2025?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** Apple Inc. issued four notes in 2025.
- **Diagnostic reviewer diagnosis:** Negative-answer specialist bad test
- **Diagnostic reviewer root cause:** evidence contains the answer; UNANSWERABLE label was wrong
- **Verifier verdict:** bad test (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 43 — `SEC\_10K\_multi\_0060\_00` · HIGH confidence

UNIT/FORMAT MISMATCH

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Precision
- **Question:** What is the amount of cash and cash equivalents as of January 25, 2026?
- ✓ **Ground truth:** 10,605
- ✗ **Model answered:** Cash and cash equivalents were \$10.605 billion as of January 25, 2026.
- **Diagnostic reviewer diagnosis:** General diagnostic specialist false positive
- **Diagnostic reviewer root cause:** Validator failed to recognize the unit conversion: the table reports \$10,605 million, which is exactly \$10.605 billion.
- **Verifier verdict:** false positive (confidence: high)
- **Validator flags:** Unsupported number: '\$10.605' does not appear in the evidence

Failure 44 — `SEC\_10K\_multi\_0520\_01` · HIGH confidence

AMBIGUOUS QUESTION

- **Category:** Precision
- **Question:** What is the total short-term unsecured funding for 2025?
- ✓ **Ground truth:** 28,668
- ✗ **Model answered:** The total short-term unsecured funding for 2025 was **\$27.341 billion**.
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** bad test
- **Diagnostic reviewer root cause:** The model selected the December 31, 2025 point-in-time total of \$27,341 million and converted it to \$27.341 billion, while the ground truth selected the 2025 average balance of \$28,668 million.
- **Verifier verdict:** bad test (confidence: high)
- **Validator flags:** Unsupported number: '\$27.341' does not appear in the evidence

Failure 45 — `SEC\_10K\_multi\_0503\_01` · HIGH confidence

UNIT/FORMAT MISMATCH

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Cross-Reference
- **Question:** What is the net capital of J.P. Morgan Securities as of December 31, 2025, and how does it compare to the minimum requirements?
- ✓ **Ground truth:** \$27,196 million, which is greater than the minimum of \$6,559 million.
- ✗ **Model answered:** As of December 31, 2025, J.P. Morgan Securities had net capital of \$27.196 billion, compared with the minimum requirement of \$6.559 billion-\$20.637 billion above the minimum.
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** false positive
- **Diagnostic reviewer root cause:** Validator failed to recognize the unit conversion from \$6,559 million to \$6.559 billion; the model also correctly calculated that \$27.196 billion is \$20.637 billion above the minimum.
- **Verifier verdict:** false positive (confidence: high)
- **Validator flags:** Unsupported number: '\$6.559' does not appear in the evidence

Failure 46 — `SEC\_10K\_multi\_1588\_01` · HIGH confidence

WRONG CALC/QUOTE

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Cross-Reference
- **Question:** What is the total revenue from concentrate operations in the United States for the year ended December 31, 2025, and how does it compare to the total revenue from finished product operations in the same period?
- ✓ **Ground truth:** The total revenue from concentrate operations in the United States for the year ended December 31, 2025, is \$8,956 million, and the total revenue from finished product operations is \$10,171 million.
- ✗ **Model answered:** For 2025, U.S. concentrate operations generated \$8,956 million, compared with \$10,171 million from finished product operations. Finished product revenue was \$1,215 million higher.
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** false positive
- **Diagnostic reviewer root cause:** The model correctly quoted both revenue figures and added a valid comparison by subtracting \$8,956 million from \$10,171 million, yielding \$1,215 million. The validator only checked whether the derived number appeared verbatim in the evidence and failed to recognize simple arithmetic.
- **Verifier verdict:** false positive (confidence: high)

- **Validator flags:** Unsupported number: '\$1,215' does not appear in the evidence

Failure 47 — `SEC\_10K\_multi\_1566\_01` · HIGH confidence

WRONG CALC/QUOTE

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Cross-Reference
- **Question:** How did the operating margin in 2025 compare to the operating margin in 2024 for the EMEA operating segment?
- ✓ **Ground truth:** The operating margin in 2025 was 39.7%, while in 2024 it was 41.4%. This indicates a decrease in operating margin for the EMEA operating segment from 2024 to 2025.
- ✗ **Model answered:** EMEA's operating margin decreased to 39.7% in 2025 from 41.4% in 2024, a decline of 1.7 percentage points.
- **Diagnostic reviewer diagnosis:** General diagnostic specialist false positive
- **Diagnostic reviewer root cause:** The model correctly quoted both operating margins and computed the 1.7-percentage-point decline by subtracting 39.7% from 41.4%. The validator only checked for literal numeric support and failed to recognize this valid derived difference.
- **Verifier verdict:** false positive (confidence: high)
- **Validator flags:** Unsupported number: '1.7' does not appear in the evidence

Failure 48 — `SEC\_10K\_multi\_1286\_00` · HIGH confidence

AMBIGUOUS QUESTION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Cross-Reference
- **Question:** What is the total amortization expense for the years ended December 31, 2025 and 2024, and how does it relate to the net carrying values of other intangible assets?
- ✓ **Ground truth:** The total amortization expense was \$292 million for 2025 and \$339 million for 2024, which relates to the net carrying values of \$1.3 billion and \$1.7 billion, respectively.
- ✗ **Model answered:** Total amortization expense was \$631 million-\$292 million in 2025 and \$339 million in 2024. Net other intangible assets were \$2.6 billion in 2025 and \$2.9 billion in 2024, comprising finite-lived assets subject to amortization and indefinite-lived assets not subject to amortization.
- **Diagnostic reviewer diagnosis:** General diagnostic specialist model error
- **Diagnostic reviewer root cause:** [verifier] The model incorrectly summed the amortization expenses for both years and failed to specify that the expense relates only to the finite-lived intangible assets, instead providing the total for all intangible assets; these are model reasoning failures, not a flaw in the test itself.
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Unsupported number: '\$631' does not appear in the evidence

Failure 49 — `SEC\_10K\_multi\_1156\_01` · HIGH confidence

AMBIGUOUS QUESTION

- **Category:** Precision
- **Question:** What is the total retained loans for noninvestment-grade risk ratings as of December 31, 2024?
- ✓ **Ground truth:** 162,432
- ✗ **Model answered:** As of December 31, 2024, noninvestment-grade retained loans totaled \$48.152 billion for loans secured by real estate and \$95.928 billion for commercial and industrial loans.
- **Diagnostic reviewer diagnosis:** General diagnostic specialist bad test

- **Diagnostic reviewer root cause:** The model interpreted the question literally and reported the noninvestment-grade totals from the two 2024 loan-category tables: \$48,152 million for secured by real estate and \$95,928 million for commercial and industrial. The expected answer instead selected the overall 'Total retained loans' figure of \$162,432 million from the secured-by-real-estate table. The validator also failed to recognize that \$48.152 billion equals \$48,152 million.
- **Verifier verdict:** bad test (confidence: high)
- **Validator flags:** Unsupported number: '\$48.152' does not appear in the evidence

Failure 50 — `SEC\_10K\_multi\_1029\_01` · HIGH confidence

UNIT/FORMAT MISMATCH

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Precision
- **Question:** What is the total investment banking fee for 2024?
- ✓ **Ground truth:** 8,910
- ✗ **Model answered:** The total investment banking fees for 2024 were **\$8.910 billion**.
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** false positive
- **Diagnostic reviewer root cause:** The model correctly converted the table's 8,910 million dollars to \$8.910 billion, but the validator performed a literal number match and did not recognize the unit conversion.
- **Verifier verdict:** false positive (confidence: high)
- **Validator flags:** Unsupported number: '\$8.910' does not appear in the evidence

Failure 51 — `SEC\_10K\_multi\_0583\_01` · HIGH confidence

BAD QUESTION

- **Category:** Cross-Reference
- **Question:** What is the total amount of derivative receivables, net of collateral, and what percentage of total derivative receivables is that?
- ✓ **Ground truth:** 11,515, 75%
- ✗ **Model answered:** Derivative receivables net of collateral totaled **\$32.807 billion**, representing approximately **53.8%** of total derivative receivables of \$60.967 billion.
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** bad test
- **Diagnostic reviewer root cause:** The model selected the actual total net derivative receivables of \$32,807 million and computed its share of gross derivative receivables ($\$32,807 / \$60,967 = 53.8\%$), whereas the ground truth combines the 1-year-or-less amount (\$11,515 million) with the investment-grade percentage (75%). The validator also failed to recognize \$32.807 billion as equivalent to \$32,807 million.
- **Verifier verdict:** bad test (confidence: high)
- **Validator flags:** Unsupported number: '\$32.807' does not appear in the evidence

Failure 52 — `SEC\_10K\_multi\_1255\_00` · HIGH confidence

BAD QUESTION

- **Category:** Precision
- **Question:** What is the total proceeds received from loan sales in 2023?
- ✓ **Ground truth:** 19,389
- ✗ **Model answered:** Total proceeds received from loan sales in 2023 were **\$19,689 million**.
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** bad test
- **Diagnostic reviewer root cause:** The model summed the 2023 cash proceeds (\$300 million) and securities proceeds (\$19,389 million) to compute total proceeds of \$19,689 million; the validator expected only the securities component and required the answer to appear verbatim in the evidence.

- **Verifier verdict:** bad test (confidence: high)
- **Validator flags:** Unsupported number: '\$19,689' does not appear in the evidence

Failure 53 — `SEC\_10K\_multi\_neg\_0080\_01` · HIGH confidence

BAD QUESTION

- **Category:** Negative (hallucination traps)
- **Question:** What is the estimated range of loss for the legal proceedings described in the passage?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** The possible loss or range of loss cannot be reasonably estimated at this time.
- **Diagnostic reviewer diagnosis:** **Negative-answer specialist** bad test
- **Diagnostic reviewer root cause:** evidence contains the answer; UNANSWERABLE label was wrong
- **Verifier verdict:** bad test (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 54 — `SEC\_10K\_multi\_1584\_00` · HIGH confidence

MATH ERROR

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Cross-Reference
- **Question:** What is the total amount of cash and cash equivalents at the end of 2025, considering the cash and cash equivalents at the beginning of the year and the net increase (decrease) in cash, cash equivalents, restricted cash and restricted cash equivalents during the year?
- ✓ **Ground truth:** 10,270
- ✗ **Model answered:** At the end of 2025, cash, cash equivalents, restricted cash and restricted cash equivalents totaled **\$11.010 billion** (\$11.488 billion – \$0.478 billion).
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** model error
- **Diagnostic reviewer root cause:** The model computed the combined year-end balance of cash, cash equivalents, restricted cash, and restricted cash equivalents ($11,488 - 478 = 11,010$) but stopped before subtracting the \$740 million of restricted cash. It therefore conflated the aggregate balance with cash and cash equivalents alone, which was \$10,270 million.
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Unsupported number: '\$11.488' does not appear in the evidence

Failure 55 — `SEC\_10K\_multi\_1403\_00` · — confidence

DETERMINISTIC:GT NUMBER CONFLICT

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Cross-Reference
- **Question:** What is the total number of derivatives qualifying as guarantees, and how does it compare to the unsettled resale and securities borrowed agreements?
- ✓ **Ground truth:** 1,168, which is less than 136,841.
- ✗ **Model answered:** Derivatives qualifying as guarantees total 49,031, compared with 137,072 for unsettled resale and securities borrowed agreements-88,041 lower.
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** model error
- **Diagnostic reviewer root cause:** [code:gt_number_conflict] Short numeric GT; response asserts a different quantity
- **Verifier verdict:** — (confidence: —)
- **Validator flags:** Unsupported number: '88,041' does not appear in the evidence

Failure 56 — `SEC\_10K\_multi\_1454\_00` · — confidence

DETERMINISTIC:GT NUMBER CONFLICT

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Precision
- **Question:** What is the total liabilities?
- ✓ **Ground truth:** 254,633
- ✗ **Model answered:** Total liabilities are **\$347,606** for Col 1 and **\$324,856** for Col 2.
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** model error
- **Diagnostic reviewer root cause:** [code:gt_number_conflict] Short numeric GT; response asserts a different quantity
- **Verifier verdict:** — (confidence: —)
- **Validator flags:** Unsupported number: '\$347,606' does not appear in the evidence

Failure 57 — `SEC\_10K\_multi\_neg\_0767\_03` · — confidence

DETERMINISTIC:ANSWERED UNANSWERABLE

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What is the volatility of natural gas commodity forward?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** The commodity volatility values are 2%, 36%, and 6%.
- **Diagnostic reviewer diagnosis:** **Negative-answer specialist** model error
- **Diagnostic reviewer root cause:** [code:answered_unanswerable] GT is UNANSWERABLE but the response asserts a concrete answer
- **Verifier verdict:** — (confidence: —)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 58 — `SEC\_10K\_multi\_0084\_01` · HIGH confidence

WRONG CALC/QUOTE

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Cross-Reference
- **Question:** What is the total revenue for the United States and Taiwan for fiscal year 2026, and how does it compare to the total revenue for the same regions in fiscal year 2025?
- ✓ **Ground truth:** The total revenue for the United States for fiscal year 2026 is \$149,617 million, and for Taiwan is \$42,345 million. For fiscal year 2025, the total revenue for the United States is \$77,482 million, and for Taiwan is \$23,600 million.
- ✗ **Model answered:** United States and Taiwan revenue totaled \$191.962 billion in fiscal 2026, versus \$101.082 billion in fiscal 2025—an increase of \$90.880 billion.
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** false positive
- **Diagnostic reviewer root cause:** Validator failed to recognize correctly computed sums and unit conversion: $\$149,617\text{M} + \$42,345\text{M} = \$191.962\text{B}$, while $\$77,482\text{M} + \$23,600\text{M} = \$101.082\text{B}$, an increase of \$90.880B.
- **Verifier verdict:** false positive (confidence: high)
- **Validator flags:** Unsupported number: '\$191.962' does not appear in the evidence

Failure 59 — `SEC\_10K\_multi\_neg\_1265\_04` · — confidence

DETERMINISTIC:ANSWERED UNANSWERABLE

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What is the total amount of goodwill allocated to the Corporate segment in 2025 that was not allocated in 2024?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** The Corporate segment's goodwill increased by **\$30 million** in 2025, from \$692 million to \$722 million.
- **Diagnostic reviewer diagnosis:** **Negative-answer specialist** model error
- **Diagnostic reviewer root cause:** [code:answered_unanswerable] GT is UNANSWERABLE but the response asserts a concrete answer
- **Verifier verdict:** — (confidence: —)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer; Unsupported number: '\$30' does not appear in the evidence

Failure 60 — `SEC\_10K\_multi\_neg\_0034\_03` · HIGH confidence

BAD QUESTION

- **Category:** Negative (hallucination traps)
- **Question:** How many notes were issued by Apple Inc. in 2023?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** Apple Inc. issued five notes in 2023.
- **Diagnostic reviewer diagnosis:** **Negative-answer specialist** bad test
- **Diagnostic reviewer root cause:** evidence contains the answer; UNANSWERABLE label was wrong
- **Verifier verdict:** bad test (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 61 — `SEC\_10K\_multi\_neg\_1609\_04` · HIGH confidence

BAD QUESTION

- **Category:** Negative (hallucination traps)
- **Question:** What is the name of the trustee for the notes issued by the Company?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** The trustee is Deutsche Bank Trust Company Americas.
- **Diagnostic reviewer diagnosis:** **Negative-answer specialist** bad test
- **Diagnostic reviewer root cause:** evidence contains the answer; UNANSWERABLE label was wrong
- **Verifier verdict:** bad test (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 62 — `SEC\_10K\_multi\_1158\_00` · HIGH confidence

WRONG CALC/QUOTE

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Cross-Reference
- **Question:** What is the total amount of retained loans for the year 2022, and how does it compare to the total retained loans for the year 2023?

- ✓ **Ground truth:** The total retained loans for 2022 is \$19,190 million, and for 2023 is \$24,287 million.
- ✗ **Model answered:** Total retained loans were \$19,190 million in 2022, compared with \$24,287 million in 2023—an increase of \$5,097 million in 2023.
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** false positive
- **Diagnostic reviewer root cause:** The model correctly quoted both retained-loan totals and added the accurately calculated difference (\$24,287 million minus \$19,190 million = \$5,097 million). The validator only checked for numbers explicitly appearing in the evidence and could not recognize the derived arithmetic.
- **Verifier verdict:** false positive (confidence: high)
- **Validator flags:** Unsupported number: '\$5,097' does not appear in the evidence

Failure 63 — `SEC\_10K\_multi\_1568\_01` · HIGH confidence

WRONG CALC/QUOTE

- **Category:** Cross-Reference
- **Question:** What is the estimated potential aggregate remaining incremental tax and interest liability for the tax years 2010 through 2025, and how does it compare to the total amount of unused backup lines of credit?
- ✓ **Ground truth:** \$14 billion
- ✗ **Model answered:** The estimated liability is approximately \$14 billion, compared with \$6.2 billion in unused backup credit lines—\$7.8 billion higher, or about 2.3 times the credit lines.
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** bad test
- **Diagnostic reviewer root cause:** Validator required numbers to appear verbatim in the evidence and could not accept the correctly derived comparison: \$14 billion minus \$6.2 billion is \$7.8 billion, and \$14 billion divided by \$6.2 billion is about 2.3.
- **Verifier verdict:** bad test (confidence: high)
- **Validator flags:** Unsupported number: '2.3' does not appear in the evidence

Failure 64 — `SEC\_10K\_multi\_0139\_00` · — confidence

DETERMINISTIC:GT NUMBER CONFLICT

- **Severity:** HIGH — an incorrect quantitative result
- **Category:** Precision
- **Question:** What is the total for New York City, New York?
- ✓ **Ground truth:** 7.1
- ✗ **Model answered:** The total for New York City, New York is 9.1.
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** model error
- **Diagnostic reviewer root cause:** [code:gt_number_conflict] Short numeric GT; response asserts a different quantity
- **Verifier verdict:** — (confidence: —)
- **Validator flags:** (none)

Failure 65 — `SEC\_10K\_multi\_0029\_01` · HIGH confidence

UNIT/FORMAT MISMATCH

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Precision
- **Question:** What is the total net sales for the Americas in 2025?
- ✓ **Ground truth:** 178,353
- ✗ **Model answered:** The Americas' total net sales in 2025 were **\$178.353 billion**.
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** false positive

- **Diagnostic reviewer root cause:** The model correctly converted the table's \$178,353 million figure to \$178.353 billion, but the validator performed a literal number match and did not account for equivalent unit conversion.
- **Verifier verdict:** false positive (confidence: high)
- **Validator flags:** Unsupported number: '\$178.353' does not appear in the evidence

Failure 66 — `SEC\_10K\_multi\_0328\_01` · — confidence

DETERMINISTIC:GT NUMBER CONFLICT

- **Severity:** HIGH — an incorrect quantitative result
- **Category:** Exception
- **Question:** What does not apply to the net charge-offs?
- ✓ **Ground truth:** \$617 million
- ✗ **Model answered:** The \$1.1 billion amount does not apply to net charge-offs; it was the net addition to the allowance for credit losses.
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** model error
- **Diagnostic reviewer root cause:** [code:gt_number_conflict] Short numeric GT; response asserts a different quantity
- **Verifier verdict:** — (confidence: —)
- **Validator flags:** (none)

Failure 67 — `SEC\_10K\_multi\_1591\_00` · HIGH confidence

WRONG CALC/QUOTE

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Cross-Reference
- **Question:** What is the total notional value of derivatives related to the Company's foreign currency economic hedges as of December 31, 2025, and how does it compare to the value as of December 31, 2024?
- ✓ **Ground truth:** The total notional value of derivatives related to the Company's foreign currency economic hedges was \$9,744 million as of December 31, 2025, and \$8,620 million as of December 31, 2024.
- ✗ **Model answered:** The total notional value was \$9,744 million as of December 31, 2025, compared with \$8,620 million in 2024—an increase of \$1,124 million.
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** false positive
- **Diagnostic reviewer root cause:** The model correctly quoted both notional values and additionally calculated the year-over-year increase (\$9,744 million minus \$8,620 million = \$1,124 million). The validator only checked for numbers explicitly present in the evidence and could not validate the derived difference.
- **Verifier verdict:** false positive (confidence: high)
- **Validator flags:** Unsupported number: '\$1,124' does not appear in the evidence

Failure 68 — `SEC\_10K\_multi\_1320\_00` · — confidence

DETERMINISTIC:GT NUMBER CONFLICT

- **Severity:** HIGH — an incorrect quantitative result
- **Category:** Precision
- **Question:** What is the fixed rate for Series DD?
- ✓ **Ground truth:** 169,625
- ✗ **Model answered:** The fixed rate for Series DD is **5.750%**.
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** model error

- **Diagnostic reviewer root cause:** [code:gt_number_conflict] Short numeric GT; response asserts a different quantity
- **Verifier verdict:** — (confidence: —)
- **Validator flags:** (none)

Failure 69 — `SEC\_10K\_multi\_1419\_01` · HIGH confidence

UNIT/FORMAT MISMATCH

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Cross-Reference
- **Question:** What is the maximum exposure of stable value contracts with contractually limited exposure, and how does it compare to the notional amount?
- ✓ **Ground truth:** 1,312 million, which is a substantially lower percentage of the notional amount
- ✗ **Model answered:** At December 31, 2025, maximum exposure was \$1.312 billion, compared with a \$35.462 billion notional amount-about 3.7% of notional.
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** false positive
- **Diagnostic reviewer root cause:** Validator failed to recognize the unit conversion: \$1.312 billion is equivalent to \$1,312 million. The comparison to the \$35.462 billion notional amount and calculated 3.7% also correctly conveys that exposure is substantially lower.
- **Verifier verdict:** false positive (confidence: high)
- **Validator flags:** Unsupported number: '\$1.312' does not appear in the evidence

Failure 70 — `SEC\_10K\_multi\_0080\_01` · HIGH confidence

UNIT/FORMAT MISMATCH

- **Category:** Cross-Reference
- **Question:** What is the income tax expense for the year ended January 25, 2026, and how does it relate to the U.S. federal statutory tax rate?
- ✓ **Ground truth:** The income tax expense for the year ended January 25, 2026, is \$21,383, which is \$8,257 less than the \$29,704 computed by applying the U.S. federal statutory tax rate of 21.0% to income before income taxes.
- ✗ **Model answered:** For the year ended January 25, 2026, income tax expense was \$21.383 billion, representing an effective tax rate of 15.1%, compared with the U.S. federal statutory rate of 21.0%.
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** bad test
- **Diagnostic reviewer root cause:** [verifier] The ground truth contains a mathematical error; the difference between the statutory tax amount (\$29,704 million) and the income tax expense (\$21,383 million) is \$8,321 million, not \$8,257 million as the ground truth claims.
- **Verifier verdict:** bad test (confidence: high)
- **Validator flags:** Unsupported number: '\$21.383' does not appear in the evidence

Failure 71 — `SEC\_10K\_multi\_neg\_0781\_01` · HIGH confidence

BAD QUESTION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** How many U.S. government agency securities are listed?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** 488 U.S. GSE and government agency securities are listed.

- **Diagnostic reviewer diagnosis:** **Negative-answer specialist** model error
- **Diagnostic reviewer root cause:** [verifier] The evidence lists a monetary value of '\$ 488' for 'U.S. GSEs and government agencies', but the question asks 'How many' securities there are, which is a count; the model incorrectly conflated the value with the count.
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 72 — `SEC\_10K\_multi\_0086\_01` · HIGH confidence

UNIT/FORMAT MISMATCH

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Cause & Effect
- **Question:** What is the result of operating leases for our offices and data centers?
- ✓ **Ground truth:** Our lease obligations primarily consist of operating leases.
- ✗ **Model answered:** As of January 25, 2026, office and data center operating leases resulted in \$3.624 billion of future minimum obligations, with a present value of \$2.944 billion and long-term liabilities of \$2.572 billion.
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** false positive
- **Diagnostic reviewer root cause:** Validator failed to recognize the unit conversion: \$3.624 billion is equivalent to the evidence's \$3,624 million. The response correctly describes the operating lease obligations for offices and data centers, while adding supported financial details.
- **Verifier verdict:** false positive (confidence: high)
- **Validator flags:** Unsupported number: '\$3.624' does not appear in the evidence

Failure 73 — `SEC\_10K\_multi\_neg\_0744\_01` · — confidence

DETERMINISTIC:ANSWERED UNANSWERABLE

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What is the total amount of derivative netting adjustments?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** The total derivative netting adjustments were \$(550,409) million.
- **Diagnostic reviewer diagnosis:** **Negative-answer specialist** model error
- **Diagnostic reviewer root cause:** [code:answered_unanswerable] GT is UNANSWERABLE but the response asserts a concrete answer
- **Verifier verdict:** — (confidence: —)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 74 — `SEC\_10K\_multi\_neg\_1135\_01` · — confidence

DETERMINISTIC:ANSWERED UNANSWERABLE

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What is the average FICO score for all retained loans?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** The weighted-average FICO scores were 775 in Col. 1 and 774 in Col. 2.
- **Diagnostic reviewer diagnosis:** **Negative-answer specialist** model error

- **Diagnostic reviewer root cause:** [code:answered_unanswerable] GT is UNANSWERABLE but the response asserts a concrete answer
- **Verifier verdict:** — (confidence: —)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 75 — `SEC\_10K\_multi\_neg\_0744\_04` · HIGH confidence

BAD QUESTION

- **Category:** Negative (hallucination traps)
- **Question:** What is the total amount of beneficial interests issued by consolidated VIEs at Level 1 of the fair value hierarchy?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** No amount was reported at Level 1 (shown as "-").
- **Diagnostic reviewer diagnosis:** **Negative-answer specialist** bad test
- **Diagnostic reviewer root cause:** evidence contains the answer; UNANSWERABLE label was wrong
- **Verifier verdict:** bad test (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 76 — `SEC\_10K\_multi\_0023\_01` · — confidence

DETERMINISTIC:GT NUMBER CONFLICT

- **Severity:** HIGH — an incorrect quantitative result
- **Category:** Precision
- **Question:** What is the value of the total?
- ✓ **Ground truth:** 132,420
- ✗ **Model answered:** The total is \$134,711.
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** model error
- **Diagnostic reviewer root cause:** [code:gt_number_conflict] Short numeric GT; response asserts a different quantity
- **Verifier verdict:** — (confidence: —)
- **Validator flags:** (none)

Failure 77 — `SEC\_10K\_multi\_1062\_00` · HIGH confidence

WRONG CALC/QUOTE

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Cross-Reference
- **Question:** What is the expected long-term rate of return on plan assets for the year ended December 31, 2025, and how does it compare to the rate for 2024?
- ✓ **Ground truth:** 5.44% for 2025 and 6.15% for 2024.
- ✗ **Model answered:** The expected long-term rate of return was 5.44% in 2025, down from 6.15% in 2024-a decrease of 0.71 percentage points.
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** false positive
- **Diagnostic reviewer root cause:** The model correctly quoted both rates and additionally computed their difference (6.15% minus 5.44% = 0.71 percentage points). The validator only checked for numbers explicitly present in the evidence and could not recognize the derived comparison.
- **Verifier verdict:** false positive (confidence: high)
- **Validator flags:** Unsupported number: '0.71' does not appear in the evidence

Failure 78 — `SEC\_10K\_multi\_neg\_0744\_02` · — confidence

DETERMINISTIC:ANSWERED UNANSWERABLE

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What is the total amount of long-term debt measured at fair value on a recurring basis?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** The total long-term debt measured at fair value on a recurring basis was **\$134,559 million**.
- **Diagnostic reviewer diagnosis:** **Negative-answer specialist** model error
- **Diagnostic reviewer root cause:** [code:answered_unanswerable] GT is UNANSWERABLE but the response asserts a concrete answer
- **Verifier verdict:** — (confidence: —)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 79 — `SEC\_10K\_multi\_neg\_0034\_01` · — confidence

DETERMINISTIC:ANSWERED UNANSWERABLE

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What is the average interest rate of the notes issued by Apple Inc. in 2025?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** The average interest rate of Apple's four notes issued in 2025 was 4.3625%.
- **Diagnostic reviewer diagnosis:** **Negative-answer specialist** model error
- **Diagnostic reviewer root cause:** [code:answered_unanswerable] GT is UNANSWERABLE but the response asserts a concrete answer
- **Verifier verdict:** — (confidence: —)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 80 — `SEC\_10K\_multi\_1471\_00` · HIGH confidence

WRONG CALC/QUOTE

- **Category:** Cross-Reference
- **Question:** What is the total value of lease right-of-use assets and the related tax-oriented investment liabilities?
- ✓ **Ground truth:** \$683 million and \$669 million
- ✗ **Model answered:** The total is **\$1.352 billion** (\$683 million + \$669 million).
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** bad test
- **Diagnostic reviewer root cause:** The model interpreted “total value” as requiring addition and correctly summed the two stated figures: \$683 million + \$669 million = \$1.352 billion. The validator expected the component values verbatim and rejected the derived total because “\$1.352” did not literally appear in the evidence.
- **Verifier verdict:** bad test (confidence: high)
- **Validator flags:** Unsupported number: '\$1.352' does not appear in the evidence

Failure 81 — `SEC\_10K\_multi\_1290\_01` · — confidence

DETERMINISTIC:DIRECTION CONFLICT

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Conditional
- **Question:** If the lease term is 7 years and the estimated useful life of leasehold improvements is 10 years, does the Firm depreciate leasehold improvements over the lease term?
- ✓ **Ground truth:** No
- ✗ **Model answered:** Yes. The Firm uses the lesser of the remaining lease term or estimated useful life, so it would depreciate the improvements over the 7-year lease term.
- **Diagnostic reviewer diagnosis:** General diagnostic specialist model error
- **Diagnostic reviewer root cause:** [code:direction_conflict] GT is 'no' but response leads 'yes'
- **Verifier verdict:** — (confidence: —)
- **Validator flags:** Direction mismatch: GT negative, response affirmative

Failure 82 — `SEC\_10K\_multi\_1587\_00` · HIGH confidence

UNIT/FORMAT MISMATCH

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Cross-Reference
- **Question:** What is the total amount of proceeds from disposals of businesses, equity method investments, and nonmarketable securities in 2025, and how does it compare to the total in 2024?
- ✓ **Ground truth:** The total amount of proceeds in 2025 is \$3,567 million, which is \$82 million more than the total in 2024, which was \$3,485 million.
- ✗ **Model answered:** Proceeds totaled \$3.567 billion in 2025, which was \$82 million higher than the \$3.485 billion reported in 2024.
- **Diagnostic reviewer diagnosis:** General diagnostic specialist false positive
- **Diagnostic reviewer root cause:** Validator failed to recognize equivalent unit conversion: \$3.567 billion equals the evidence's \$3,567 million. The model's 2024 amount and \$82 million comparison also match the correct answer.
- **Verifier verdict:** false positive (confidence: high)
- **Validator flags:** Unsupported number: '\$3.567' does not appear in the evidence

Failure 83 — `SEC\_10K\_multi\_0075\_01` · HIGH confidence

UNIT/FORMAT MISMATCH

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Precision
- **Question:** What was the total of other long-term liabilities as of January 26, 2025?
- ✓ **Ground truth:** 4,245
- ✗ **Model answered:** The total other long-term liabilities as of January 26, 2025 was \$4.245 billion.
- **Diagnostic reviewer diagnosis:** General diagnostic specialist false positive
- **Diagnostic reviewer root cause:** Validator failed to recognize unit conversion: the table reports \$4,245 million, which is equivalent to \$4.245 billion.
- **Verifier verdict:** false positive (confidence: high)
- **Validator flags:** Unsupported number: '\$4.245' does not appear in the evidence

Failure 84 — `SEC\_10K\_multi\_0019\_01` · HIGH confidence

UNIT/FORMAT MISMATCH

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Cross-Reference
- **Question:** What is the total net sales for the years ended September 27, 2025, September 28, 2024, and September 30, 2023, and how do they compare to each other?
- ✓ **Ground truth:** \$416,161, \$391,035, and \$383,285
- ✗ **Model answered:** Total net sales were \$416.161 billion in 2025, \$391.035 billion in 2024, and \$383.285 billion in 2023. Sales increased by \$7.750 billion from 2023 to 2024 and by \$25.126 billion from 2024 to 2025.
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** false positive
- **Diagnostic reviewer root cause:** Validator failed to account for unit conversion: the table reports amounts in millions, so \$416,161 million, \$391,035 million, and \$383,285 million equal \$416.161 billion, \$391.035 billion, and \$383.285 billion. The stated year-over-year increases are also correct.
- **Verifier verdict:** false positive (confidence: high)
- **Validator flags:** Unsupported number: '\$416.161' does not appear in the evidence

Failure 85 — `SEC\_10K\_multi\_1160\_01` · — confidence

DETERMINISTIC:GT NUMBER CONFLICT

- **Severity:** HIGH — an incorrect quantitative result
- **Category:** Cross-Reference
- **Question:** What is the total retained loans amount in the U.S. and the total loan delinquency amount in the 30-89 days past due category, respectively?
- ✓ **Ground truth:** \$165,545 and 918
- ✗ **Model answered:** The total retained loans in the U.S. were **\$626,060**, and loans 30-89 days past due totaled **\$2,913**.
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** model error
- **Diagnostic reviewer root cause:** [code:gt_number_conflict] Short numeric GT; response asserts a different quantity
- **Verifier verdict:** — (confidence: —)
- **Validator flags:** (none)

Failure 86 — `SEC\_10K\_multi\_0084\_00` · HIGH confidence

MISINTERPRETED

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Cross-Reference
- **Question:** What is the total depreciation and amortization expense for the Compute & Networking segment for fiscal years 2026 and 2025, and how does it compare to the Graphics segment for the same years?
- ✓ **Ground truth:** The total depreciation and amortization expense for the Compute & Networking segment for fiscal years 2026 and 2025 is \$1.6 billion and \$732 million, respectively. For the Graphics segment, the total depreciation and amortization expense for fiscal years 2026 and 2025 is \$590 million and \$372 million, respectively.
- ✗ **Model answered:** Compute & Networking's total depreciation and amortization expense was \$2.332 billion for fiscal 2026 and 2025, compared with \$962 million for Graphics-\$1.370 billion higher.
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** model error

- **Diagnostic reviewer root cause:** The model summed the two fiscal-year figures for each segment (\$1.6B + \$732M = \$2.332B; \$590M + \$372M = \$962M) instead of reporting the separate 2026 and 2025 amounts requested. Its aggregate comparison is arithmetically valid but omits the year-by-year answer.
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Unsupported number: '\$2.332' does not appear in the evidence

Failure 87 — `SEC\_10K\_multi\_0851\_00` · HIGH confidence

AMBIGUOUS QUESTION

- **Category:** Precision
- **Question:** What is the total liabilities measured at fair value on a nonrecurring basis?
- ✓ **Ground truth:** 5
- ✗ **Model answered:** At December 31, 2024, total liabilities measured at fair value on a nonrecurring basis were \$0 million.
- **Diagnostic reviewer diagnosis:** General diagnostic specialist bad test
- **Diagnostic reviewer root cause:** The model selected the December 31, 2024 table, where dashes indicate zero liabilities, while the ground truth uses the preceding table's total of \$5 million.
- **Verifier verdict:** bad test (confidence: high)
- **Validator flags:** Unsupported number: '\$0' does not appear in the evidence

Failure 88 — `SEC\_10K\_multi\_0019\_00` · HIGH confidence

WRONG CALC/QUOTE

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Cross-Reference
- **Question:** What is the maximum one-day loss in fair value of the Company's foreign currency derivative positions due to adverse movements in exchange rates, as of September 27, 2025, and how does it compare to the loss in 2024?
- ✓ **Ground truth:** \$590 million and \$538 million
- ✗ **Model answered:** The maximum one-day loss was \$590 million as of September 27, 2025, compared with \$538 million in 2024—an increase of \$52 million.
- **Diagnostic reviewer diagnosis:** General diagnostic specialist false positive
- **Diagnostic reviewer root cause:** The response correctly quoted the 2025 and 2024 losses and additionally computed the increase as \$590 million minus \$538 million = \$52 million. The validator only accepted numbers explicitly stated in the evidence and failed to recognize this valid arithmetic derivation.
- **Verifier verdict:** false positive (confidence: high)
- **Validator flags:** Unsupported number: '\$52' does not appear in the evidence

Failure 89 — `SEC\_10K\_multi\_1299\_01` · HIGH confidence

BAD QUESTION

- **Category:** Precision
- **Question:** What is the total senior debt at December 31, 2025?
- ✓ **Ground truth:** 226,071
- ✗ **Model answered:** Total senior debt at December 31, 2025 was **\$234.532 billion** (\$226.071 billion fixed-rate plus \$8.461 billion variable-rate).
- **Diagnostic reviewer diagnosis:** General diagnostic specialist bad test
- **Diagnostic reviewer root cause:** The model summed the fixed-rate and variable-rate senior debt totals (\$226.071 billion + \$8.461 billion), while the reference answer incorrectly quotes only the fixed-rate senior debt row.

- **Verifier verdict:** bad test (confidence: high)
- **Validator flags:** Unsupported number: '\$234.532' does not appear in the evidence

Failure 90 — `SEC\_10K\_multi\_neg\_1265\_03` · — confidence

DETERMINISTIC:ANSWERED UNANSWERABLE

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What is the total amount of goodwill allocated to the Asset & Wealth Management segment in 2024 that was not allocated in 2023?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** There was no additional goodwill allocated in 2024; Asset & Wealth Management goodwill decreased by \$61 million, from \$8.582 billion in 2023 to \$8.521 billion in 2024.
- **Diagnostic reviewer diagnosis:** **Negative-answer specialist** model error
- **Diagnostic reviewer root cause:** [code:answered_unanswerable] GT is UNANSWERABLE but the response asserts a concrete answer
- **Verifier verdict:** — (confidence: —)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer; Unsupported number: '\$61' does not appear in the evidence

Failure 91 — `SEC\_10K\_multi\_1159\_01` · — confidence

DETERMINISTIC:GT NUMBER CONFLICT

- **Severity:** HIGH — an incorrect quantitative result
- **Category:** Precision
- **Question:** What is the percentage of criticized to total retained loans secured by real estate for 2025?
- ✓ **Ground truth:** 4.43
- ✗ **Model answered:** The percentage was 6.37% in 2025.
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** model error
- **Diagnostic reviewer root cause:** [code:gt_number_conflict] Short numeric GT; response asserts a different quantity
- **Verifier verdict:** — (confidence: —)
- **Validator flags:** (none)

Failure 92 — `SEC\_10K\_multi\_0910\_00` · — confidence

DETERMINISTIC:GT NUMBER CONFLICT

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Cross-Reference
- **Question:** What is the total amount for Asset Managers and Consumer & Retail combined?
- ✓ **Ground truth:** 302,693
- ✗ **Model answered:** The combined total for Asset Managers and Consumer & Retail is 286,793.
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** model error
- **Diagnostic reviewer root cause:** [code:gt_number_conflict] Short numeric GT; response asserts a different quantity
- **Verifier verdict:** — (confidence: —)
- **Validator flags:** Unsupported number: '286,793' does not appear in the evidence

Failure 93 — `SEC\_10K\_multi\_1584\_01` · HIGH confidence

UNIT/FORMAT MISMATCH

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Cross-Reference
- **Question:** What is the total amount of equity attributable to shareowners of The Coca-Cola Company at the end of 2025, considering the equity attributable to shareowners at the beginning of the year and the net income attributable to shareowners of The Coca-Cola Company?
- ✓ **Ground truth:** 32,169
- ✗ **Model answered:** At the end of 2025, total equity attributable to shareowners of The Coca-Cola Company was \$32.169 billion.
- **Diagnostic reviewer diagnosis:** General diagnostic specialist false positive
- **Diagnostic reviewer root cause:** The model converted the evidence's \$32,169 million into \$32.169 billion, but the validator performed a literal number match and could not recognize the unit conversion.
- **Verifier verdict:** false positive (confidence: high)
- **Validator flags:** Unsupported number: '\$32.169' does not appear in the evidence

Failure 94 — `SEC\_10K\_multi\_0275\_00` · HIGH confidence

BAD QUESTION

- **Category:** Calculation
- **Question:** What is TCE for December 31, 2025?
- ✓ **Ground truth:** 289,002
- ✗ **Model answered:** Insufficient information
- **Diagnostic reviewer diagnosis:** Calculation specialist bad test
- **Diagnostic reviewer root cause:** model matches evidence formula; GT is wrong
- **Verifier verdict:** bad test (confidence: high)
- **Validator flags:** Number mismatch: expected 289,002, not found in response

Failure 95 — `SEC\_10K\_multi\_1549\_00` · HIGH confidence

WRONG CALC/QUOTE

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Cross-Reference
- **Question:** What is the estimated increase in net periodic pension cost if the discount rate decreases by 50 basis points, and how does this compare to the estimated increase if the expected long-term rate of return on plan assets decreases by 50 basis points?
- ✓ **Ground truth:** \$9 million and \$19 million, respectively
- ✗ **Model answered:** A 50-basis-point decrease in the discount rate would increase 2026 net periodic pension cost by \$9 million, compared with a \$19 million increase from the same decrease in the expected long-term return on plan assets-\$10 million more.
- **Diagnostic reviewer diagnosis:** General diagnostic specialist false positive
- **Diagnostic reviewer root cause:** The model correctly extracted the \$9 million and \$19 million increases and additionally computed their \$10 million difference. The validator only checked for numbers explicitly present in the evidence and failed to recognize the valid subtraction (\$19 million minus \$9 million).
- **Verifier verdict:** false positive (confidence: high)
- **Validator flags:** Unsupported number: '\$10' does not appear in the evidence

Failure 96 — `SEC\_10K\_multi\_1589\_00` · HIGH confidence

MATH ERROR

- **Severity:** HIGH — an incorrect quantitative result
- **Category:** Calculation
- **Question:** What is the total fair value of the Company's equity securities as of December 31, 2025?
- ✓ **Ground truth:** 2589
- ✗ **Model answered:** 2633
- **Diagnostic reviewer diagnosis:** **Calculation specialist** model error
- **Diagnostic reviewer root cause:** model computed a different number than the formula yields
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Number mismatch: expected 2589, not found in response

Failure 97 — `SEC\_10K\_multi\_neg\_1169\_04` · — confidence

DETERMINISTIC:ANSWERED UNANSWERABLE

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What is the total amount of payment deferrals provided in 2023?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** The total amount of payment deferrals provided in 2023 was \$8 million.
- **Diagnostic reviewer diagnosis:** **Negative-answer specialist** model error
- **Diagnostic reviewer root cause:** [code:answered_unanswerable] GT is UNANSWERABLE but the response asserts a concrete answer
- **Verifier verdict:** — (confidence: —)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer; Unsupported number: '\$8' does not appear in the evidence

Failure 98 — `SEC\_10K\_multi\_neg\_0516\_00` · HIGH confidence

BAD QUESTION

- **Category:** Negative (hallucination traps)
- **Question:** What is the average interest rate on non-interest-bearing deposits at JPMorgan Chase's U.S. offices as of December 31, 2025?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** Not applicable (NA); non-interest-bearing deposits do not have an average interest rate.
- **Diagnostic reviewer diagnosis:** **Negative-answer specialist** bad test
- **Diagnostic reviewer root cause:** evidence contains the answer; UNANSWERABLE label was wrong
- **Verifier verdict:** bad test (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 99 — `SEC\_10K\_multi\_1173\_00` · HIGH confidence

AMBIGUOUS QUESTION

- **Category:** Cross-Reference
- **Question:** What is the percentage of loan modifications to total retained Commercial and industrial loans, and how does it compare to the number of single modifications?

- ✓ **Ground truth:** 0.71% of loan modifications to total retained Commercial and industrial loans, with 1,180 single modifications.
- ✗ **Model answered:** Loan modifications totaled 1.12% of retained commercial and industrial loans. Single modifications accounted for \$1.644 billion, comprising 0.99% of retained loans.
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** bad test
- **Diagnostic reviewer root cause:** The model aggregated all listed modification categories to get 1.12% and combined the two single-modification rows to get \$1.644 billion and 0.99%, whereas the expected answer selected only the single-modification term-extension row (0.71% and \$1.180 billion).
- **Verifier verdict:** bad test (confidence: high)
- **Validator flags:** Unsupported number: '1.12%' does not appear in the evidence

Failure 100 — `SEC\_10K\_multi\_0516\_00` · HIGH confidence

BAD QUESTION

- **Category:** Cross-Reference
- **Question:** What is the total amount of available cash and securities as of December 31, 2025, and how does it compare to the total amount as of December 31, 2024?
- ✓ **Ground truth:** \$1.5 trillion and \$1.4 trillion, respectively, with a decrease of \$100 billion compared to the previous year.
- ✗ **Model answered:** The Firm had approximately \$1.5 trillion of available cash and securities as of December 31, 2025, up \$0.1 trillion from \$1.4 trillion as of December 31, 2024.
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** bad test
- **Diagnostic reviewer root cause:** The model correctly computed the year-over-year change from the stated totals: \$1.5 trillion minus \$1.4 trillion equals a \$0.1 trillion (\$100 billion) increase. The validator treated the derived \$0.1 trillion as unsupported because it was not quoted verbatim, while the ground truth incorrectly labels the change as a decrease.
- **Verifier verdict:** bad test (confidence: high)
- **Validator flags:** Unsupported number: '\$0.1' does not appear in the evidence

Failure 101 — `SEC\_10K\_multi\_0854\_00` · HIGH confidence

BAD QUESTION

- **Category:** Cross-Reference
- **Question:** What were the total net gains/losses from the measurement alternative for the years ended December 31, 2023, 2024, and 2025?
- ✓ **Ground truth:** \$197 million, \$232 million, and \$122 million, respectively
- ✗ **Model answered:** The total was a net loss of **\$307 million**: losses of \$232 million in 2023 and \$197 million in 2024, offset by a \$122 million gain in 2025.
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** bad test
- **Diagnostic reviewer root cause:** The model correctly reversed the evidence's reverse-chronological presentation to assign a \$232 million loss to 2023, a \$197 million loss to 2024, and a \$122 million gain to 2025, then summed them to a net \$307 million loss. The validator rejected the derived total because it was not stated verbatim, while the ground truth swaps the 2023 and 2024 figures.
- **Verifier verdict:** bad test (confidence: high)
- **Validator flags:** Unsupported number: '\$307' does not appear in the evidence

Failure 102 — `SEC\_10K\_multi\_0630\_01` · — confidence

DETERMINISTIC:GT NUMBER CONFLICT

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream

- **Category:** Precision
- **Question:** What is the total for loans due after one year at fixed interest rates?
- ✓ **Ground truth:** 14,356
- ✗ **Model answered:** The total loans due after one year at fixed interest rates were **\$86,224**.
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** model error
- **Diagnostic reviewer root cause:** [code:gt_number_conflict] Short numeric GT; response asserts a different quantity
- **Verifier verdict:** — (confidence: —)
- **Validator flags:** Unsupported number: '\$86,224' does not appear in the evidence

Failure 103 — `SEC\_10K\_multi\_neg\_1609\_01` · — confidence

DETERMINISTIC:ANSWERED UNANSWERABLE

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What is the average interest rate of the notes issued by the Company in 2020?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** The average interest rate of the 15 notes issued in 2020 was approximately **2.15%**.
- **Diagnostic reviewer diagnosis:** **Negative-answer specialist** model error
- **Diagnostic reviewer root cause:** [code:answered_unanswerable] GT is UNANSWERABLE but the response asserts a concrete answer
- **Verifier verdict:** — (confidence: —)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer; Unsupported number: '2.15%' does not appear in the evidence

Failure 104 — `SEC\_10K\_multi\_neg\_0581\_02` · — confidence

DETERMINISTIC:ANSWERED UNANSWERABLE

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What is the total value of derivative receivables that are not collateralized?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** The total derivative receivables not collateralized were **\$28,886 million**.
- **Diagnostic reviewer diagnosis:** **Negative-answer specialist** model error
- **Diagnostic reviewer root cause:** [code:answered_unanswerable] GT is UNANSWERABLE but the response asserts a concrete answer
- **Verifier verdict:** — (confidence: —)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 105 — `SEC\_10K\_multi\_1335\_01` · HIGH confidence

BAD QUESTION

- **Category:** Calculation
- **Question:** What is the aggregate purchase price of common stock repurchases in 2023?
- ✓ **Ground truth:** 69.5
- ✗ **Model answered:** Insufficient information
- **Diagnostic reviewer diagnosis:** **Calculation specialist** bad test
- **Diagnostic reviewer root cause:** model matches evidence formula; GT is wrong
- **Verifier verdict:** bad test (confidence: high)

- **Validator flags:** Number mismatch: expected 69.5, not found in response

Failure 106 — `SEC\_10K\_multi\_0275\_01` · HIGH confidence

BAD QUESTION

- **Category:** Calculation
- **Question:** What is TBVPS for December 31, 2025?
- ✓ **Ground truth:** 0.845
- ✗ **Model answered:** Insufficient information
- **Diagnostic reviewer diagnosis:** Calculation specialist bad test
- **Diagnostic reviewer root cause:** model matches evidence formula; GT is wrong
- **Verifier verdict:** bad test (confidence: high)
- **Validator flags:** Number mismatch: expected 0.845, not found in response

Complete Failure Listing — Without VERITROOPER (Baseline)

194 confirmed model errors + 47 excluded bad tests (malformed question, not counted against the model). Each entry shows the full question, the model's full answer, the ground truth, the diagnostic reviewer's diagnosis, and (where available) the Verifier's independent audit verdict.

Failure shape at a glance — Without VERITROOPER (Baseline)

Category	Scored incorrect	Model error	Disputed (adjudicators cleared)	Bad test	Other	Review recommended
Precision	82	54	13	15	0	0
Negative (hallucination traps)	80	71	3	6	0	0
Cross-Reference	69	20	28	21	0	0
Conditional	5	4	0	1	0	0
Calculation	3	0	0	3	0	0
Cause & Effect	2	0	1	1	0	0
TOTAL	241	149	45	47	0	0

Scored incorrect is the count that reaches the headline — the deterministic check scored these answers wrong. The next four columns say what adjudication then found, and they sum to it. **Disputed** means the diagnostic reviewer (no independent Verifier ran on this audit) judged the answer correct; those answers still count against the model in the headline, because the deterministic check decides the score and the model verdicts are advisory.

Severity mix (confirmed errors): 192 HIGH, 2 MEDIUM. Severity is a domain-agnostic heuristic from the error type — a confidently-wrong answer outranks a missing one; re-weight for your risk context.

Failure 1 — `SEC\_10K\_multi\_neg\_1409\_01` · — confidence

DETERMINISTIC:ANSWERED UNANSWERABLE

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What is the percentage of risk participations for other letters of credit compared to total commitments?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** For 2025, risk participations for other letters of credit were **\$514 million**, compared with **\$4.529 billion** of total commitments:

$$\frac{514}{4.529} \times 100 \approx 11.35\%$$

Thus, risk participations were approximately **11.4% of total commitments**.

- **Diagnostic reviewer diagnosis:** **Negative-answer specialist** model error
- **Diagnostic reviewer root cause:** [code:answered_unanswerable] GT is UNANSWERABLE but the response asserts a concrete answer
- **Verifier verdict:** — (confidence: —)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer; Unsupported number: '\$4.529' does not appear in the evidence

Failure 2 — `SEC\_10K\_multi\_1345\_00` · — confidence

DETERMINISTIC:GT NUMBER CONFLICT

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Precision
- **Question:** What is the net change for DVA on fair value option elected liabilities?
- ✓ **Ground truth:** 1,025
- ✗ **Model answered:** The net change in DVA on fair value option-elected liabilities was **\$(773) million after tax in 2025 (\$1.025) billion pre-tax**.
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** model error
- **Diagnostic reviewer root cause:** [code:gt_number_conflict] Short numeric GT; response asserts a different quantity
- **Verifier verdict:** — (confidence: —)
- **Validator flags:** Unsupported number: '1.025' does not appear in the evidence

Failure 3 — `SEC\_10K\_multi\_1568\_00` · HIGH confidence

UNSUPPORTED NUMBER

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Cross-Reference
- **Question:** What is the total amount of unused backup lines of credit for general corporate purposes as of December 31, 2025, and how does it relate to the Company's capital allocation priorities?
- ✓ **Ground truth:** \$6.2 billion
- ✗ **Model answered:** As of December 31, 2025, the Company had approximately **\$6.2 billion** (\$6.15 billion) in unused backup lines of credit for general corporate purposes.

These lines support liquidity and financial flexibility alongside the Company's capital allocation priorities:

1. Investing in business operations
2. Growing dividend payments
3. Expanding its beverage portfolio through acquisitions
4. Repurchasing shares with excess cash over time

- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** model error

- **Diagnostic reviewer root cause:** [verifier] The model invented the figure '(\$6.15 billion),' which is not present in the evidence; the evidence only states the company had '\$6.2 billion in unused backup lines of credit.'

- **Verifier verdict:** model error (confidence: high)

- **Validator flags:** Unsupported number: '\$6.15' does not appear in the evidence

Failure 4 — `SEC\_10K\_multi\_neg\_0516\_01` · HIGH confidence

BAD QUESTION

- **Category:** Negative (hallucination traps)

- **Question:** What is the total amount of deposits that are covered by deposit insurance protection at JPMorgan Chase as of December 31, 2025?

- ✓ **Ground truth:** UNANSWERABLE

- ✗ **Model answered:** As of December 31, 2025, deposits covered by insurance protection totaled approximately **\$1,000.7 billion**:

\$2,559.3 billion total deposits – \$1,558.6 billion uninsured deposits = \$1,000.7 billion insured deposits.

- **Diagnostic reviewer diagnosis:** **Negative-answer specialist** bad test

- **Diagnostic reviewer root cause:** The model derived insured deposits by subtracting disclosed estimated uninsured deposits from total deposits, while the benchmark required an explicit stated figure and labeled the question unanswerable.

- **Verifier verdict:** bad test (confidence: high)

- **Validator flags:** Hallucination: model answered a question the evidence cannot answer; Unsupported number: '\$1,000.7' does not appear in the evidence

Failure 5 — `SEC\_10K\_multi\_neg\_0714\_02` · HIGH confidence

BAD QUESTION

- **Category:** Negative (hallucination traps)

- **Question:** What is the name of the specific accounting standard that the new guidance is intended to clarify?

- ✓ **Ground truth:** UNANSWERABLE

- ✗ **Model answered:** The guidance is intended to clarify **Revenue from Contracts with Customers (ASC 606)**, specifically the accounting for share-based noncash consideration from customers.

- **Diagnostic reviewer diagnosis:** **Negative-answer specialist** bad test

- **Diagnostic reviewer root cause:** [verifier] The ground truth is "UNANSWERABLE", but the evidence explicitly names the accounting standards the guidance clarifies in its title: "Derivatives and Hedging and Revenue from Contracts with Customers", making the test's expected answer incorrect.

- **Verifier verdict:** bad test (confidence: high)

- **Validator flags:** Hallucination: model answered a question the evidence cannot answer; Unsupported number: '606' does not appear in the evidence

Failure 6 — `SEC\_10K\_multi\_0064\_00` · HIGH confidence

BAD QUESTION

- **Category:** Cross-Reference
- **Question:** What is the total amount of cash and cash equivalents at the beginning of the period for the year ended January 25, 2026, and the year ended January 26, 2025?
- ✓ **Ground truth:** 10,605 and 8,589
- ✗ **Model answered:** Cash and cash equivalents at the beginning of each period were:
 - Year ended January 25, 2026: **\$8.589 billion**
 - Year ended January 26, 2025: **\$7.280 billion**

Total: \$15.869 billion.

- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** bad test
- **Diagnostic reviewer root cause:** The model correctly read the cash-flow row for beginning-of-period cash (8,589 and 7,280 million) and summed the two figures; the reference answer instead used the end-of-period balances (10,605 and 8,589 million). The validator also failed to recognize \$8.589 billion as equivalent to 8,589 million.
- **Verifier verdict:** bad test (confidence: high)
- **Validator flags:** Unsupported number: '\$8.589' does not appear in the evidence

Failure 7 — `SEC\_10K\_multi\_neg\_1585\_05` · — confidence

DETERMINISTIC:ANSWERED UNANSWERABLE

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What is the average return on investment for the Company's investments in limited partnerships?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** For the alternative-energy limited partnerships:
 - **2025 ROI:** $\$241\text{M} \div \$306\text{M} = 78.8\%$
 - **2024 ROI:** $\$323\text{M} \div \$226\text{M} = 142.9\%$

The **average annual ROI was approximately 110.8%.**

On a combined basis, total benefits of \$564M divided by \$532M invested equals **106.0%.**

- **Diagnostic reviewer diagnosis:** **Negative-answer specialist** model error
- **Diagnostic reviewer root cause:** [code:answered_unanswerable] GT is UNANSWERABLE but the response asserts a concrete answer
- **Verifier verdict:** — (confidence: —)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer; Unsupported number: '\$241M' does not appear in the evidence

Failure 8 — `SEC\_10K\_multi\_0304\_00` · HIGH confidence

MISINTERPRETED

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Precision
- **Question:** What was the value of mortgage loans 90 or more days past due and insured by U.S. government agencies at December 31, 2023?
- ✓ **Ground truth:** 123 million

- **✗ Model answered:** At December 31, 2023, mortgage loans 90 or more days past due and insured by U.S. government agencies totaled **\$182 million** (\$123 million in CCB and \$59 million in CIB).
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** model error
- **Diagnostic reviewer root cause:** The model misread the stated \$123 million total as a CCB component and added an unsupported \$59 million CIB amount, producing \$182 million instead of quoting the 2023 figure.
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Unsupported number: '\$182' does not appear in the evidence

Failure 9 — `SEC\_10K\_multi\_neg\_1062\_02` · — confidence

DETERMINISTIC:ANSWERED UNANSWERABLE

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What is the total value of the defined benefit pension and OPEB plans' liabilities as of December 31, 2025?
- **✓ Ground truth:** UNANSWERABLE
- **✗ Model answered:** As of December 31, 2025, the defined benefit pension and OPEB plans' total projected benefit obligations were **\$14.724 billion**.
- **Diagnostic reviewer diagnosis:** **Negative-answer specialist** model error
- **Diagnostic reviewer root cause:** [code:answered_unanswerable] GT is UNANSWERABLE but the response asserts a concrete answer
- **Verifier verdict:** — (confidence: —)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer; Unsupported number: '\$14.724' does not appear in the evidence

Failure 10 — `SEC\_10K\_multi\_neg\_1595\_01` · — confidence

DETERMINISTIC:ANSWERED UNANSWERABLE

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What is the average annual return on the U.S. pension plan's assets for the year 2025?
- **✓ Ground truth:** UNANSWERABLE
- **✗ Model answered:** The expected average annual return on the U.S. pension plan's assets for 2025 was **6.75%**.
- **Diagnostic reviewer diagnosis:** **Negative-answer specialist** model error
- **Diagnostic reviewer root cause:** [code:answered_unanswerable] GT is UNANSWERABLE but the response asserts a concrete answer
- **Verifier verdict:** — (confidence: —)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 11 — `SEC\_10K\_multi\_0934\_00` · HIGH confidence

UNSUPPORTED NUMBER

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Precision
- **Question:** What is the total fair value of trading assets and liabilities?
- **✓ Ground truth:** 665,587

- **Model answered:** As of December 31, 2025, the total fair value was:
- **Trading assets:** \$606.382 billion gross; **\$57.777 billion net**
- **Trading liabilities:** \$596.738 billion gross; **\$46.329 billion net**
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** model error
- **Diagnostic reviewer root cause:** The model used figures from the wrong reporting period or table and separately reported unrelated gross/net trading assets and liabilities instead of reading the stated total fair value of 665,587 from the evidence row.
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Unsupported number: '\$606.382' does not appear in the evidence

Failure 12 — `SEC\_10K\_multi\_1588\_00` · HIGH confidence

UNIT/FORMAT MISMATCH

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Cross-Reference
- **Question:** What is the total amount of assets held for sale as of December 31, 2025, and how does it compare to the total amount of liabilities held for sale?
- **Ground truth:** The total amount of assets held for sale as of December 31, 2025, is \$5,342 million, and the total amount of liabilities held for sale is \$2,570 million.
- **Model answered:** Assets held for sale totaled **\$5.342 billion** as of December 31, 2025. The evidence does not provide the complete total for liabilities held for sale, so an exact comparison cannot be determined; it only shows **\$816 million** of accounts payable and accrued expenses within those liabilities.
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** model error
- **Diagnostic reviewer root cause:** The model correctly converted \$5,342 million to \$5.342 billion, but failed to read the table's final 'Liabilities held for sale' total of \$2,570 million and instead treated the \$816 million accounts-payable component as the only available liability figure. The validator's complaint about \$5.342 is merely a unit-format parsing issue, but the response is still substantively incomplete and incorrect about liabilities.
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Unsupported number: '\$5.342' does not appear in the evidence

Failure 13 — `SEC\_10K\_multi\_1614\_01` · — confidence

DETERMINISTIC:GT NUMBER CONFLICT

- **Severity:** HIGH — an incorrect quantitative result
- **Category:** Precision
- **Question:** What is the number of the certification executed by John Murphy?
- **Ground truth:** 32.1
- **Model answered:** The certification executed by John Murphy is **Exhibit 31.2**.
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** model error
- **Diagnostic reviewer root cause:** [code:gt_number_conflict] Short numeric GT; response asserts a different quantity
- **Verifier verdict:** — (confidence: —)
- **Validator flags:** (none)

Failure 14 — `SEC\_10K\_multi\_1113\_01` · HIGH confidence

UNSUPPORTED NUMBER

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream

- **Category:** Precision
- **Question:** What is the total loan balance at December 31, 2025?
- ✓ **Ground truth:** 1,493,429
- ✗ **Model answered:** The total retained loan balance at December 31, 2025 was **\$1.409 trillion** (\$1,408,905 million).
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** model error
- **Diagnostic reviewer root cause:** Model read the 'Retained' row total (\$1,408,905 million) instead of the overall 'Total' row (\$1,493,429 million), thereby excluding held-for-sale and fair-value loans.
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Unsupported number: '\$1.409' does not appear in the evidence

Failure 15 — `SEC\_10K\_multi\_0216\_01` · HIGH confidence

BAD QUESTION

- **Category:** Cross-Reference
- **Question:** What is the cumulative total return for JPMorgan Chase & Co. common stock over the five-year period ending December 31, 2025, and how does it compare to the S&P 500 Index?
- ✓ **Ground truth:** The cumulative total return for JPMorgan Chase & Co. common stock was \$289.18, which is higher than the S&P 500 Index return of \$196.12.
- ✗ **Model answered:** Over the five years ended December 31, 2025:
- **JPMorgan Chase common stock: 189.18% cumulative total return** (\$100 grew to \$289.18)
- **S&P 500 Index: 96.12% cumulative total return** (\$100 grew to \$196.12)

JPMorgan Chase outperformed the S&P 500 by **93.06 percentage points**, assuming dividends were reinvested.

- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** bad test
- **Diagnostic reviewer root cause:** The model converted the table's indexed ending values into percentage gains by subtracting the initial \$100, while also explicitly providing the expected ending values. The validator only accepted verbatim evidence figures and treated the correctly derived 189.18% as unsupported.
- **Verifier verdict:** bad test (confidence: high)
- **Validator flags:** Unsupported number: '189.18%' does not appear in the evidence

Failure 16 — `SEC\_10K\_multi\_0583\_00` · HIGH confidence

WRONG CALC/QUOTE

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Cross-Reference
- **Question:** What is the total amount of loans retained after 5 years, and what percentage of investment-grade loans is that?
- ✓ **Ground truth:** 175,215, 68%
- ✗ **Model answered:** As of December 31, 2025, loans retained after five years totaled **\$189.819 billion**, and **68%** of total retained loans were investment-grade.
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** model error
- **Diagnostic reviewer root cause:** Model used the wrong reporting period or table value, stating a December 31, 2025 amount of \$189.819 billion instead of the evidence's December 31, 2024 after-five-years loans-retained amount of \$175.215 billion. It correctly extracted the 68% investment-grade ratio.
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Unsupported number: '\$189.819' does not appear in the evidence

Failure 17 — `SEC\_10K\_multi\_1501\_01` · HIGH confidence

UNSUPPORTED NUMBER

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Precision
- **Question:** What is the decrease in deposits with banks due to change in rate?
- ✓ **Ground truth:** (2,792)
- ✗ **Model answered:** The decrease due to changes in rates was **\$4.352 billion** in 2025 versus 2024:
 - U.S.: \$2.792 billion
 - Non-U.S.: \$1.560 billion
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** model error
- **Diagnostic reviewer root cause:** [verifier] The model hallucinated a value for 'Non-U.S.' deposits, which is not present in the evidence, and used this fabricated number to calculate its final answer; this is a clear model error regardless of any ambiguity in the question.
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Unsupported number: '\$4.352' does not appear in the evidence

Failure 18 — `SEC\_10K\_multi\_0747\_01` · HIGH confidence

BAD QUESTION

- **Category:** Precision
- **Question:** What is the total value of asset-backed securities?
- ✓ **Ground truth:** 11,316
- ✗ **Model answered:** The total fair value of available-for-sale asset-backed securities was **\$23.906 billion** as of December 31, 2025.
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** bad test
- **Diagnostic reviewer root cause:** The model appears to have pulled an asset-backed securities figure from a different table or reporting period rather than reading the provided row, which shows a total of 2,823.
- **Verifier verdict:** bad test (confidence: high)
- **Validator flags:** Unsupported number: '\$23.906' does not appear in the evidence

Failure 19 — `SEC\_10K\_multi\_neg\_0767\_01` · — confidence

DETERMINISTIC:ANSWERED UNANSWERABLE

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What is the correlation between equity and commodity derivatives?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** No direct equity-commodity correlation is disclosed. Separately, the weighted-average inputs are **54% for equity correlation** and **1% for commodity correlation**.
- **Diagnostic reviewer diagnosis:** **Negative-answer specialist** model error
- **Diagnostic reviewer root cause:** [code:answered_unanswerable] GT is UNANSWERABLE but the response asserts a concrete answer
- **Verifier verdict:** — (confidence: —)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 20 — `SEC\_10K\_multi\_0825\_01` · HIGH confidence

MISINTERPRETED

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Precision
- **Question:** What is the total number of debt instruments?
- ✓ **Ground truth:** 2,178
- ✗ **Model answered:** Total debt instruments were **\$2,182 million (\$2.182 billion)** as of December 31, 2025.
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** model error
- **Diagnostic reviewer root cause:** Model misread or mistyped the stated total debt instruments figure, reporting 2,182 million instead of the table-row total of 2,178 million.
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Unsupported number: '\$2.182' does not appear in the evidence

Failure 21 — `SEC\_10K\_multi\_1181\_00` · — confidence

DETERMINISTIC:GT NUMBER CONFLICT

- **Severity:** HIGH — an incorrect quantitative result
- **Category:** Cross-Reference
- **Question:** What is the percentage of loan modifications to total retained Other loans for single modifications?
- ✓ **Ground truth:** 0.10
- ✗ **Model answered:** For 2025, single modifications represented **0.03%** of total retained Other loans.
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** model error
- **Diagnostic reviewer root cause:** [code:gt_number_conflict] Short numeric GT; response asserts a different quantity
- **Verifier verdict:** — (confidence: —)
- **Validator flags:** (none)

Failure 22 — `SEC\_10K\_multi\_1159\_00` · HIGH confidence

MISINTERPRETED

- **Category:** Precision
- **Question:** What is the total retained Secured by real estate loans for 2025?
- ✓ **Ground truth:** 105,130
- ✗ **Model answered:** The total retained loans secured by real estate for 2025 were **\$165.545 billion**.
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** bad test
- **Diagnostic reviewer root cause:** The model correctly summed/read the 2025 total column and converted \$165,545 million to \$165.545 billion; the validator failed to recognize the unit conversion, while the reference answer misread the Multifamily value as the total.
- **Verifier verdict:** bad test (confidence: high)
- **Validator flags:** Unsupported number: '\$165.545' does not appear in the evidence

Failure 23 — `SEC\_10K\_multi\_1350\_00` · HIGH confidence

VALIDATOR OVER-FIRED

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Precision

- **Question:** What is the total income tax expense and effective tax rate for the first period?
- ✓ **Ground truth:** 15,547
- ✗ **Model answered:** The total income tax expense was **\$15.547 billion**, with an effective tax rate of **21.4%**.
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** false positive
- **Diagnostic reviewer root cause:** [verifier] The model correctly identified the income tax expense as \$15,547 and the effective rate as 21.4%; its conversion of the expense to '\$15.547 billion' is a reasonable interpretation of financial data, and the validator incorrectly flagged this correct answer.
- **Verifier verdict:** false positive (confidence: high)
- **Validator flags:** Unsupported number: '\$15.547' does not appear in the evidence

Failure 24 — `SEC\_10K\_multi\_neg\_0034\_04` · HIGH confidence

BAD QUESTION

- **Category:** Negative (hallucination traps)
- **Question:** What is the total number of notes issued by Apple Inc. in 2022?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** Apple Inc. issued **4 notes** in 2022.
- **Diagnostic reviewer diagnosis:** **Negative-answer specialist** bad test
- **Diagnostic reviewer root cause:** The model counted the four note series listed in the August 8, 2022 officer's certificate (3.250%, 3.350%, 3.950%, and 4.100% notes), interpreting them as notes issued in 2022.
- **Verifier verdict:** bad test (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 25 — `SEC\_10K\_multi\_neg\_0714\_04` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What is the name of the specific firm or organization responsible for issuing the new guidance?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** The Financial Accounting Standards Board (FASB).
- **Diagnostic reviewer diagnosis:** **Negative-answer specialist** model error
- **Diagnostic reviewer root cause:** The model inferred FASB from the accounting-guidance context rather than limiting its answer to information explicitly provided in the evidence.
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 26 — `SEC\_10K\_multi\_1115\_01` · HIGH confidence

VALIDATOR OVER-FIRED

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Cross-Reference
- **Question:** What is the total amount of sales for the year ended December 31, 2025?
- ✓ **Ground truth:** 55,583
- ✗ **Model answered:** The total amount of sales for the year ended December 31, 2025 was **\$55.583 billion** (\$55,583 million).
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** false positive
- **Diagnostic reviewer root cause:** pre-Doctor equivalence: numeric match: model has \$55,583 ~ GT 55,583
- **Verifier verdict:** false positive (confidence: high)

- **Validator flags:** Unsupported number: '\$55.583' does not appear in the evidence

Failure 27 — `SEC\_10K\_multi\_neg\_0825\_03` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What is the total value of equity securities held by individuals?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** As of December 31, 2025, exposure to **Individuals and Individual Entities** totaled **\$167.7 billion**.
- **Diagnostic reviewer diagnosis:** **Negative-answer specialist** model error
- **Diagnostic reviewer root cause:** [verifier] The model hallucinated a specific value and category ('Individuals and Individual Entities') not present in the evidence, whereas the correct action was to state the question is unanswerable, as reflected in the ground truth.
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 28 — `SEC\_10K\_multi\_0572\_00` · HIGH confidence

BAD QUESTION

- **Category:** Cross-Reference
- **Question:** What is the total amount of consumer loans excluding credit card loans, and what is the total amount of consumer loans including credit card loans?
- ✓ **Ground truth:** The total amount of consumer loans excluding credit card loans is \$402,258, and the total amount of consumer loans including credit card loans is \$247,797.
- ✗ **Model answered:** As of December 31, 2025:
 - **Consumer loans excluding credit card loans:** **\$402.258 billion**
 - **Total consumer loans including credit card loans:** **\$650.055 billion**
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** bad test
- **Diagnostic reviewer root cause:** The model read the actual 'Total consumer loans' figure of 650,055 and converted the table's presumed millions to \$650.055 billion; it likewise converted 402,258 to \$402.258 billion. The validator failed to recognize the unit conversion, while the reference incorrectly used the 'Total credit card loans' figure (247,797) as the total including credit cards.
- **Verifier verdict:** bad test (confidence: high)
- **Validator flags:** Unsupported number: '\$402.258' does not appear in the evidence

Failure 29 — `SEC\_10K\_multi\_0822\_01` · HIGH confidence

TABLE MISREAD

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Cross-Reference
- **Question:** What is the fair value at December 31, 2023, and what is the change in unrealized gains/(losses) related to financial instruments held at December 31, 2023?
- ✓ **Ground truth:** 34,564, 738
- ✗ **Model answered:** For **long-term debt**, the fair value at December 31, 2023 was **\$27.726 billion**, and the change in unrealized gains/(losses) on instruments held at that date was **\$2.870 billion**.
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** model error

- **Diagnostic reviewer root cause:** The model read the wrong table row and misaligned the columns, reporting the long-term-debt figure of \$27,726 million instead of the requested total fair value of \$34,564 million, and gave an incorrect unrealized gain rather than \$738 million.
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Unsupported number: '\$27.726' does not appear in the evidence

Failure 30 — `SEC\_10K\_multi\_0060\_01` · HIGH confidence

UNIT/FORMAT MISMATCH

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Precision
- **Question:** What is the amount of cash used in financing activities for fiscal year 2026?
- **Ground truth:** 48,474
- **Model answered:** Cash used in financing activities for fiscal year 2026 was **\$48.474 billion**.
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** false positive
- **Diagnostic reviewer root cause:** The model correctly converted the table's \$48,474 million to \$48.474 billion, but the validator performed a literal number match and could not recognize the unit conversion.
- **Verifier verdict:** false positive (confidence: high)
- **Validator flags:** Unsupported number: '\$48.474' does not appear in the evidence

Failure 31 — `SEC\_10K\_multi\_1403\_01` · HIGH confidence

UNSUPPORTED NUMBER

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Cross-Reference
- **Question:** How do the unsettled repurchase and securities loaned agreements compare to the unsettled resale and securities borrowed agreements?
- **Ground truth:** 52,308 is less than 136,841.
- **Model answered:** - **Unsettled repurchase and securities loaned agreements:** **\$52.9 billion** at year-end 2025, down from **\$67.0 billion** in 2024. At settlement, the Firm receives cash and provides securities collateral.
- **Unsettled resale and securities borrowed agreements:** **\$137.1 billion** in 2025, up from **\$115.9 billion** in 2024. At settlement, the Firm advances cash and receives securities collateral.

Thus, the resale and securities borrowed commitments were about **\$84.2 billion larger** in 2025.

- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** model error
- **Diagnostic reviewer root cause:** [verifier] The model invented information not present in the evidence by describing the settlement process for both agreements, stating for example that for one, 'the Firm receives cash and provides securities collateral.'
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Unsupported number: '\$52.9' does not appear in the evidence

Failure 32 — `SEC\_10K\_multi\_0754\_01` · HIGH confidence

UNSUPPORTED NUMBER

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Cross-Reference
- **Question:** What is the total amount of beneficial interests issued by consolidated VIEs and long-term debt combined?

- ✓ **Ground truth:** $1 + 100,780 = 100,781$
- ✗ **Model answered:** As of December 31, 2025:
- Beneficial interests issued by consolidated VIEs: **\$27.951 billion**
- Long-term debt: **\$435.206 billion**

Combined total: \$463.157 billion.

- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** model error
- **Diagnostic reviewer root cause:** The model used unrelated December 31, 2025 figures instead of reading the total values from the provided fair-value table. The relevant rows show beneficial interests of 1 and long-term debt of 100,780, totaling 100,781.
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Unsupported number: '\$435.206' does not appear in the evidence

Failure 33 — `SEC\_10K\_multi\_neg\_1169\_03` · — confidence

DETERMINISTIC:ANSWERED UNANSWERABLE

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What is the percentage of loans that were extended by more than 21 months in 2024?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** In 2024, **0.07% of retained "Other" loans** had term extensions averaging 28 months, which was more than 21 months.
- **Diagnostic reviewer diagnosis:** **Negative-answer specialist** model error
- **Diagnostic reviewer root cause:** [code:answered_unanswerable] GT is UNANSWERABLE but the response asserts a concrete answer
- **Verifier verdict:** — (confidence: —)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 34 — `SEC\_10K\_multi\_0725\_01` · HIGH confidence

UNSUPPORTED NUMBER

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Precision
- **Question:** What is the total liabilities at December 31, 2024?
- ✓ **Ground truth:** 27,777
- ✗ **Model answered:** Total liabilities at December 31, 2024 were **\$3,658,056 million** (approximately **\$3.66 trillion**).
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** model error
- **Diagnostic reviewer root cause:** Model answered with JPMorganChase's firm-wide total liabilities from a different balance-sheet context instead of reading the 2024 total-liabilities row in the provided consolidated-VIE table, which is \$27,777 million.
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Unsupported number: '\$3,658,056' does not appear in the evidence

Failure 35 — `SEC\_10K\_multi\_neg\_1602\_01` · — confidence

DETERMINISTIC:ANSWERED UNANSWERABLE

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream

- **Category:** Negative (hallucination traps)
- **Question:** What is the average annual return on the Company's pension plan assets?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** The Company's expected long-term annual return on pension plan assets for 2025 was 6.75%.
- **Diagnostic reviewer diagnosis:** Negative-answer specialist model error
- **Diagnostic reviewer root cause:** [code:answered_unanswerable] GT is UNANSWERABLE but the response asserts a concrete answer
- **Verifier verdict:** — (confidence: —)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 36 — `SEC\_10K\_multi\_1115\_00` · HIGH confidence

BAD QUESTION

- **Category:** Cross-Reference
- **Question:** What is the total amount of retained loans reclassified to held-for-sale for the year ended December 31, 2025?
- ✓ **Ground truth:** 55,583
- ✗ **Model answered:** The total retained loans reclassified to held-for-sale in 2025 was **\$1.552 billion**.
- **Diagnostic reviewer diagnosis:** General diagnostic specialist bad test
- **Diagnostic reviewer root cause:** The model selected an unsupported figure, likely from another row or disclosure, rather than the retained-loans reclassification row.
- **Verifier verdict:** bad test (confidence: high)
- **Validator flags:** Unsupported number: '\$1.552' does not appear in the evidence

Failure 37 — `SEC\_10K\_multi\_neg\_1095\_02` · — confidence

DETERMINISTIC:ANSWERED UNANSWERABLE

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What is the average credit risk exposure for resale and securities borrowed agreements as of December 31, 2025?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** The average net credit risk exposure was **\$31.967 billion**, calculated as:
 - Resale agreements: **\$12.209 billion**
 - Securities borrowed: **\$51.725 billion**

$$\begin{aligned} & \backslash [\\ & (\$12.209\text{B} + \$51.725\text{B}) / 2 = \$31.967\text{B} \\ & \backslash] \end{aligned}$$

- **Diagnostic reviewer diagnosis:** Negative-answer specialist model error
- **Diagnostic reviewer root cause:** [code:answered_unanswerable] GT is UNANSWERABLE but the response asserts a concrete answer
- **Verifier verdict:** — (confidence: —)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer; Unsupported number: '\$31.967' does not appear in the evidence

Failure 38 — `SEC\_10K\_multi\_0072\_01` · HIGH confidence

VALIDATOR OVER-FIRED

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Precision
- **Question:** What was the total inventories as of January 26, 2025?
- ✓ **Ground truth:** 10,080
- ✗ **Model answered:** Total inventories as of January 26, 2025 were **\$10.080 billion**.
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** false positive
- **Diagnostic reviewer root cause:** Validator failed to recognize unit equivalence: the evidence reports 10,080 in millions, which equals \$10.080 billion.
- **Verifier verdict:** false positive (confidence: high)
- **Validator flags:** Unsupported number: '\$10.080' does not appear in the evidence

Failure 39 — `SEC\_10K\_multi\_1241\_01` · HIGH confidence

UNSUPPORTED NUMBER

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Precision
- **Question:** What is the amount of amortization losses from investments in 2023?
- ✓ **Ground truth:** zero
- ✗ **Model answered:** Amortization losses from investments were **\$1.561 billion in 2023**.
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** model error
- **Diagnostic reviewer root cause:** Model selected an unrelated or hallucinated figure instead of mapping the respectively ordered values for 2025, 2024, and 2023; the 2023 amortization loss is zero.
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Unsupported number: '\$1.561' does not appear in the evidence

Failure 40 — `SEC\_10K\_multi\_0318\_01` · HIGH confidence

BAD QUESTION

- **Category:** Cross-Reference
- **Question:** How does the bank's revenue from card income compare to its total net revenue in 2025?
- ✓ **Ground truth:** The bank's revenue from card income was \$2,436 million, which is part of the total net revenue of \$53,766 million in 2025.
- ✗ **Model answered:** In 2025, Consumer & Community Banking generated **\$2.23 billion in card income**, compared with **\$76.03 billion in total net revenue**. Card income represented approximately **2.9%** of total net revenue.
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** bad test
- **Diagnostic reviewer root cause:** The model used figures from the wrong business segment/table and appears to have mistaken the 2023 card-income figure of about \$2.23 billion for 2025, rather than reading the 2025 Commercial & Investment Bank row.
- **Verifier verdict:** bad test (confidence: high)
- **Validator flags:** Unsupported number: '\$2.23' does not appear in the evidence

Failure 41 — `SEC\_10K\_multi\_0725\_00` · HIGH confidence

UNIT/FORMAT MISMATCH

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Precision
- **Question:** What is the total assets at December 31, 2025?
- ✓ **Ground truth:** 43,295
- ✗ **Model answered:** Total assets at December 31, 2025 were **\$4,424.9 billion**.
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** model error
- **Diagnostic reviewer root cause:** Model selected JPMorganChase's firmwide total assets from a different balance-sheet context rather than the total assets of consolidated VIEs in the provided table. The evidence reports 43,295 million (\$43.295 billion), not \$4,424.9 billion.
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Unsupported number: '\$4,424.9' does not appear in the evidence

Failure 42 — `SEC\_10K\_multi\_neg\_0852\_00` · — confidence

DETERMINISTIC:ANSWERED UNANSWERABLE

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What is the total value of level 3 assets measured at fair value on a recurring basis as of December 31, 2025?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** The total value was **\$25.107 billion** as of December 31, 2025.
- **Diagnostic reviewer diagnosis:** **Negative-answer specialist** model error
- **Diagnostic reviewer root cause:** [code:answered_unanswerable] GT is UNANSWERABLE but the response asserts a concrete answer
- **Verifier verdict:** — (confidence: —)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 43 — `SEC\_10K\_multi\_neg\_1255\_04` · — confidence

DETERMINISTIC:ANSWERED UNANSWERABLE

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What is the total number of loans sold to the U.S. GSEs in 2025?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** In 2025, the carrying value of loans sold to U.S. GSEs was **\$30.496 billion**.
- **Diagnostic reviewer diagnosis:** **Negative-answer specialist** model error
- **Diagnostic reviewer root cause:** [code:answered_unanswerable] GT is UNANSWERABLE but the response asserts a concrete answer
- **Verifier verdict:** — (confidence: —)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer; Unsupported number: '\$30.496' does not appear in the evidence

Failure 44 — `SEC\_10K\_multi\_1125\_00` · HIGH confidence

UNSUPPORTED NUMBER

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Precision
- **Question:** What is the total retained loans for 2024?
- ✓ **Ground truth:** 376,334
- ✗ **Model answered:** Total retained loans for 2024 were **\$690.396 billion (\$690,396 million)**.
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** model error
- **Diagnostic reviewer root cause:** Model selected or computed a figure from the wrong loan category/table instead of reading the 2024 'Total retained loans' row, which explicitly reports \$376,334 million.
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Unsupported number: '\$690.396' does not appear in the evidence

Failure 45 — `SEC\_10K\_multi\_neg\_1169\_02` · — confidence

DETERMINISTIC:ANSWERED UNANSWERABLE

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What is the total number of loans modified in 2023?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** In 2023, total retained credit card loan modifications were **\$648 million** in amortized cost basis.
- **Diagnostic reviewer diagnosis:** **Negative-answer specialist** model error
- **Diagnostic reviewer root cause:** [code:answered_unanswerable] GT is UNANSWERABLE but the response asserts a concrete answer
- **Verifier verdict:** — (confidence: —)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 46 — `SEC\_10K\_multi\_0744\_01` · HIGH confidence

BAD QUESTION

- **Category:** Precision
- **Question:** What is the total liabilities measured at fair value on a recurring basis?
- ✓ **Ground truth:** 550,409
- ✗ **Model answered:** Total liabilities measured at fair value on a recurring basis were **\$586.175 billion** as of December 31, 2025.
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** bad test
- **Diagnostic reviewer root cause:** The model produced an unsupported aggregate of \$586.175 billion rather than reading the total from the table's final column.
- **Verifier verdict:** bad test (confidence: high)
- **Validator flags:** Unsupported number: '\$586.175' does not appear in the evidence

Failure 47 — `SEC\_10K\_multi\_0494\_01` · HIGH confidence

VALIDATOR OVER-FIRED

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream

- **Category:** Precision
- **Question:** What was the Tier 1 capital as of December 31, 2025?
- ✓ **Ground truth:** 307,630
- ✗ **Model answered:** Tier 1 capital was **\$307.63 billion** as of December 31, 2025.
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** false positive
- **Diagnostic reviewer root cause:** Validator failed to account for unit conversion: the table reports \$307,630 million, which is exactly \$307.63 billion.
- **Verifier verdict:** false positive (confidence: high)
- **Validator flags:** Unsupported number: '\$307.63' does not appear in the evidence

Failure 48 — `SEC\_10K\_multi\_1341\_01` · HIGH confidence

VALIDATOR OVER-FIRED

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Precision
- **Question:** What is the net unrealized gain/(loss) arising during the period for 2023?
- ✓ **Ground truth:** 3,891
- ✗ **Model answered:** For 2023, the net unrealized gain arising during the period was **\$3.891 billion pre-tax**, or **\$2.969 billion after tax**.
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** false positive
- **Diagnostic reviewer root cause:** Validator failed to recognize a unit conversion: the evidence reports 3,891 million, which is equivalent to \$3.891 billion. The model also correctly included the after-tax equivalent of 2,969 million (\$2.969 billion).
- **Verifier verdict:** false positive (confidence: high)
- **Validator flags:** Unsupported number: '\$3.891' does not appear in the evidence

Failure 49 — `SEC\_10K\_multi\_0080\_00` · HIGH confidence

VALIDATOR OVER-FIRED

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Cross-Reference
- **Question:** What is the total amount of current income taxes for the year ended January 25, 2026, and how does it compare to the year ended January 26, 2025?
- ✓ **Ground truth:** The total current income taxes for the year ended January 25, 2026, is \$22,807, which is \$7,184 more than the \$15,623 for the year ended January 26, 2025.
- ✗ **Model answered:** Total current income taxes were **\$22.807 billion** for the year ended January 25, 2026, compared with **\$15.623 billion** for the year ended January 26, 2025. This was an **increase of \$7.184 billion**, or approximately **46%**.
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** false positive
- **Diagnostic reviewer root cause:** The model correctly converted the table's figures from millions (\$22,807 million and \$15,623 million) to billions (\$22.807 billion and \$15.623 billion); the validator failed to recognize the unit conversion because the decimal-form number does not appear verbatim in the evidence.
- **Verifier verdict:** false positive (confidence: high)
- **Validator flags:** Unsupported number: '\$15.623' does not appear in the evidence

Failure 50 — `SEC\_10K\_multi\_neg\_0767\_00` · — confidence

DETERMINISTIC:ANSWERED UNANSWERABLE

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What is the average price of oil in BBL?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** The average oil price is **\$202 per barrel (BBL)**.
- **Diagnostic reviewer diagnosis:** **Negative-answer specialist** model error
- **Diagnostic reviewer root cause:** [code:answered_unanswerable] GT is UNANSWERABLE but the response asserts a concrete answer
- **Verifier verdict:** — (confidence: —)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 51 — `SEC\_10K\_multi\_1155\_01` · HIGH confidence

UNIT/FORMAT MISMATCH

- **Category:** Cross-Reference
- **Question:** What is the total noninvestment-grade loans for 2025 and 2024, and what percentage of total retained loans are criticized nonaccrual loans for each year?
- ✓ **Ground truth:** The total noninvestment-grade loans for 2025 is \$46,670 and for 2024 is \$48,152. The percentage of total retained loans that are criticized nonaccrual loans for 2025 is 1.01% and for 2024 is 0.89%.
- ✗ **Model answered:** - **2025:** Total noninvestment-grade loans were **\$251.0 billion**; criticized nonaccrual loans were **0.56%** of total retained loans.
- **2024:** Total noninvestment-grade loans were **\$218.7 billion**; criticized nonaccrual loans were **0.57%** of total retained loans.
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** bad test
- **Diagnostic reviewer root cause:** The model selected the rightmost aggregate column pair (251,003 and 218,726, with 0.56% and 0.57%) and converted the amounts from millions to billions, while the expected answer uses the first column pair (46,670 and 48,152, with 1.01% and 0.89%).
- **Verifier verdict:** bad test (confidence: high)
- **Validator flags:** Unsupported number: '\$218.7' does not appear in the evidence

Failure 52 — `SEC\_10K\_multi\_neg\_0034\_02` · — confidence

DETERMINISTIC:ANSWERED UNANSWERABLE

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What is the total value of the notes issued by Apple Inc. in 2025?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** Apple issued **\$4.5 billion** in notes in 2025.
- **Diagnostic reviewer diagnosis:** **Negative-answer specialist** model error
- **Diagnostic reviewer root cause:** [code:answered_unanswerable] GT is UNANSWERABLE but the response asserts a concrete answer
- **Verifier verdict:** — (confidence: —)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer; Unsupported number: '\$4.5' does not appear in the evidence

Failure 53 — `SEC\_10K\_multi\_0520\_00` · HIGH confidence

UNIT/FORMAT MISMATCH

- **Category:** Precision
- **Question:** What is the total short-term unsecured funding for 2024?
- ✓ **Ground truth:** 24,985
- ✗ **Model answered:** The total short-term unsecured funding for 2024 was **\$28.517 billion**.
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** bad test
- **Diagnostic reviewer root cause:** Model selected the 2024 year-end balance (\$28,517 million) rather than the 2024 average balance (\$24,985 million); its \$28.517 billion figure is supported by the table and is unit-equivalent to \$28,517 million.
- **Verifier verdict:** bad test (confidence: high)
- **Validator flags:** Unsupported number: '\$28.517' does not appear in the evidence

Failure 54 — `SEC\_10K\_multi\_1409\_01` · — confidence

DETERMINISTIC:GT NUMBER CONFLICT

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Cross-Reference
- **Question:** What was the total amount for other letters of credit as of December 31, 2025?
- ✓ **Ground truth:** \$514 million
- ✗ **Model answered:** The total contractual amount for other letters of credit was **\$4.529 billion** as of December 31, 2025.
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** model error
- **Diagnostic reviewer root cause:** [code:gt_number_conflict] Short numeric GT; response asserts a different quantity
- **Verifier verdict:** — (confidence: —)
- **Validator flags:** Unsupported number: '\$4.529' does not appear in the evidence

Failure 55 — `SEC\_10K\_multi\_neg\_1595\_03` · — confidence

DETERMINISTIC:ANSWERED UNANSWERABLE

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What is the total value of the Company's other postretirement benefit plans' assets as of December 31, 2025?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** As of December 31, 2025, the other postretirement benefit plans' assets totaled **\$67 million**.
- **Diagnostic reviewer diagnosis:** **Negative-answer specialist** model error
- **Diagnostic reviewer root cause:** [code:answered_unanswerable] GT is UNANSWERABLE but the response asserts a concrete answer
- **Verifier verdict:** — (confidence: —)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 56 — `SEC\_10K\_multi\_0368\_00` · HIGH confidence

VALIDATOR OVER-FIRED

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Cross-Reference
- **Question:** What were the total assets under custody in Fixed Income and Equity Markets as of December 31, 2025?
- ✓ **Ground truth:** 18,322 billion and 17,954 billion
- ✗ **Model answered:** As of December 31, 2025, assets under custody were:
 - **Fixed Income:** \$18.322 trillion
 - **Equity:** \$17.954 trillion
 - **Combined total:** \$36.276 trillion
- **Diagnostic reviewer diagnosis:** General diagnostic specialist false positive
- **Diagnostic reviewer root cause:** Validator failed to recognize equivalent unit conversion: 18,322 billion equals 18.322 trillion, and 17,954 billion equals 17.954 trillion. The model also correctly summed them to 36.276 trillion.
- **Verifier verdict:** false positive (confidence: high)
- **Validator flags:** Unsupported number: '\$18.322' does not appear in the evidence

Failure 57 — `SEC\_10K\_multi\_0320\_01` · — confidence

DETERMINISTIC:GT NUMBER CONFLICT

- **Severity:** HIGH — an incorrect quantitative result
- **Category:** Cross-Reference
- **Question:** What was the return on equity for the years ended December 31, 2023, 2024, and 2025?
- ✓ **Ground truth:** 14%, 18%, and 18%
- ✗ **Model answered:** - **2023:** 38%
 - **2024:** 32%
 - **2025:** 32%
- **Diagnostic reviewer diagnosis:** General diagnostic specialist model error
- **Diagnostic reviewer root cause:** [code:gt_number_conflict] Short numeric GT; response asserts a different quantity
- **Verifier verdict:** — (confidence: —)
- **Validator flags:** (none)

Failure 58 — `SEC\_10K\_multi\_1455\_01` · HIGH confidence

UNIT/FORMAT MISMATCH

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Precision
- **Question:** What is the net cash provided by/(used in) operating activities for 2024?
- ✓ **Ground truth:** -21,981
- ✗ **Model answered:** Net cash used in operating activities for 2024 was \$42.012 billion (reported as \$(42,012) million).
- **Diagnostic reviewer diagnosis:** General diagnostic specialist model error
- **Diagnostic reviewer root cause:** Model selected or derived the wrong cash-flow figure instead of reading the 2024 operating-activities total from the table; the stated \$42.012 billion does not match the reported \$(21,981) million.

- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Unsupported number: '\$42.012' does not appear in the evidence

Failure 59 — `SEC\_10K\_multi\_0551\_00` · HIGH confidence

TABLE MISREAD

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Precision
- **Question:** What was the total nonaccrual loans at December 31, 2025?
- ✓ **Ground truth:** 9,857
- ✗ **Model answered:** Total nonaccrual loans at December 31, 2025 were **\$5.159 billion**.
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** model error
- **Diagnostic reviewer root cause:** Model extracted or hallucinated a figure from the wrong table or portfolio category instead of reading the Firmwide total nonaccrual loans row. The evidence reports \$9,857 million (\$9.857 billion), not \$5.159 billion.
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Unsupported number: '\$5.159' does not appear in the evidence

Failure 60 — `SEC\_10K\_multi\_0055\_00` · HIGH confidence

WRONG PERIOD

- **Category:** Cross-Reference
- **Question:** What is the total value of shares withheld for tax withholding obligations in fiscal year 2026, and how does it compare to the total value of shares repurchased during the same period?
- ✓ **Ground truth:** 51 million shares for \$7.9 billion, which is less than the \$1.5 billion spent on repurchasing 8 million shares.
- ✗ **Model answered:** In fiscal 2026, NVIDIA withheld shares valued at **\$7.948 billion** for tax obligations, compared with **\$40.388 billion** of shares repurchased. Thus, tax-withheld shares were **\$32.44 billion lower**, or about **20%** of the repurchase value.
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** bad test
- **Diagnostic reviewer root cause:** The model interpreted “during the same period” as the full fiscal year and used more precise fiscal-year figures, whereas the reference answer uses the \$1.5 billion repurchase occurring after fiscal 2026 ended.
- **Verifier verdict:** bad test (confidence: high)
- **Validator flags:** Unsupported number: '\$7.948' does not appear in the evidence

Failure 61 — `SEC\_10K\_multi\_0942\_00` · HIGH confidence

UNIT/FORMAT MISMATCH

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Conditional
- **Question:** If the derivative receivables are not nettable on the Consolidated balance sheets, what is the total amount?
- ✓ **Ground truth:** 57,777
- ✗ **Model answered:** As of December 31, 2025, collateral related to derivative receivables not nettable on the Consolidated balance sheets totaled **\$28.891 billion**.
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** model error

- **Diagnostic reviewer root cause:** [verifier] The model provided the value for "Collateral not nettable on the Consolidated balance sheets" (\$28,891 million), but the question explicitly asked for the amount of "derivative receivables", making the model's answer incorrect as it conflated two distinct items from the evidence.
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Unsupported number: '\$28.891' does not appear in the evidence

Failure 62 — `SEC\_10K\_multi\_neg\_0034\_08` · — confidence

DETERMINISTIC:ANSWERED UNANSWERABLE

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What is the average interest rate of the notes issued by Apple Inc. in 2022?
- **Ground truth:** UNANSWERABLE
- **Model answered:** The average interest rate of Apple's four notes issued in 2022 was **3.6625%**, or approximately **3.66%**.

\[

$$(3.250\% + 3.350\% + 3.950\% + 4.100\%) \div 4 = 3.6625\%$$

 \]

- **Diagnostic reviewer diagnosis:** **Negative-answer specialist** model error
- **Diagnostic reviewer root cause:** [code:answered_unanswerable] GT is UNANSWERABLE but the response asserts a concrete answer
- **Verifier verdict:** — (confidence: —)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer; Unsupported number: '3.6625%' does not appear in the evidence

Failure 63 — `SEC\_10K\_multi\_0747\_00` · HIGH confidence

BAD QUESTION

- **Category:** Precision
- **Question:** What is the total value of equity securities?
- **Ground truth:** 417,428
- **Model answered:** As of December 31, 2025, total equity securities were **\$2.633 billion** (\$2.589 billion at fair value plus \$44 million under the measurement alternative).
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** bad test
- **Diagnostic reviewer root cause:** The model pulled an unsupported figure from a different table or context rather than reading the equity securities row in the supplied evidence.
- **Verifier verdict:** bad test (confidence: high)
- **Validator flags:** Unsupported number: '\$2.633' does not appear in the evidence

Failure 64 — `SEC\_10K\_multi\_1354\_00` · HIGH confidence

BAD QUESTION

- **Category:** Cross-Reference
- **Question:** What was the total income tax expense for the year ended December 31, 2025, and what portion of it was due to tax benefits?
- **Ground truth:** 15,547, 9,936
- **Model answered:** Total income tax expense was **\$15.547 billion** in 2025. It included **\$629 million of tax benefits** from resolved tax audits, equivalent to approximately **4.0%** of the total expense.
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** bad test

- **Diagnostic reviewer root cause:** The model correctly converted \$15,547 million to \$15.547 billion and used the explicitly stated \$629 million in tax benefits, while the reference answer appears to have mistakenly selected \$9,936 million from the 'Total current income tax expense' row. The validator also failed to recognize the million-to-billion unit conversion.
- **Verifier verdict:** bad test (confidence: high)
- **Validator flags:** Unsupported number: '\$15.547' does not appear in the evidence

Failure 65 — `SEC\_10K\_multi\_neg\_1274\_04` · — confidence

DETERMINISTIC:ANSWERED UNANSWERABLE

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What is the total amount of contractual service fees included in income for the year?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** Contractual service fees, late fees, and other ancillary fees included in income totaled \$1.635 billion for 2025.
- **Diagnostic reviewer diagnosis:** Negative-answer specialist model error
- **Diagnostic reviewer root cause:** [code:answered_unanswerable] GT is UNANSWERABLE but the response asserts a concrete answer
- **Verifier verdict:** — (confidence: —)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer; Unsupported number: '\$1.635' does not appear in the evidence

Failure 66 — `SEC\_10K\_multi\_0020\_00` · — confidence

DETERMINISTIC:GT NUMBER CONFLICT

- **Severity:** HIGH — an incorrect quantitative result
- **Category:** Precision
- **Question:** What is the threshold for a fiscal year?
- ✓ **Ground truth:** 52 or 53 weeks
- ✗ **Model answered:** The threshold is 10% or more of total revenue in a fiscal year.
- **Diagnostic reviewer diagnosis:** General diagnostic specialist model error
- **Diagnostic reviewer root cause:** [code:gt_number_conflict] Short numeric GT; response asserts a different quantity
- **Verifier verdict:** — (confidence: —)
- **Validator flags:** (none)

Failure 67 — `SEC\_10K\_multi\_0518\_01` · HIGH confidence

UNSUPPORTED NUMBER

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Cross-Reference
- **Question:** What is the total deposits in U.S. offices in the second year, given the total deposits in the first year and the growth rate?
- ✓ **Ground truth:** 1,918,734
- ✗ **Model answered:** Total deposits in U.S. offices in the second year were \$1,991,661 million (about \$1.992 trillion), up approximately 3.8% from \$1,918,734 million.
- **Diagnostic reviewer diagnosis:** General diagnostic specialist model error

- **Diagnostic reviewer root cause:** [verifier] The model answered with the value for the third year (1,991,661) when the question asked for the second year, which is correctly identified in the evidence as 1,918,734.
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Unsupported number: '3.8%' does not appear in the evidence

Failure 68 — `SEC\_10K\_multi\_0388\_01` · — confidence

DETERMINISTIC:GT NUMBER CONFLICT

- **Severity:** HIGH — an incorrect quantitative result
- **Category:** Precision
- **Question:** What is the allowance for loan losses to period-end loans?
- ✓ **Ground truth:** 0.20
- ✗ **Model answered:** The allowance for loan losses to period-end loans retained was **1.37% in 2025**.
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** model error
- **Diagnostic reviewer root cause:** [code:gt_number_conflict] Short numeric GT; response asserts a different quantity
- **Verifier verdict:** — (confidence: —)
- **Validator flags:** (none)

Failure 69 — `SEC\_10K\_multi\_0518\_00` · HIGH confidence

VALIDATOR OVER-FIRED

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Cross-Reference
- **Question:** What is the total interest-bearing deposits in non-U.S. offices in the third year, given the total interest-bearing deposits in the first year and the growth rate?
- ✓ **Ground truth:** 408,419
- ✗ **Model answered:** The total interest-bearing deposits in non-U.S. offices in the third year (2023) were **\$408,419 million**, or approximately **\$408.4 billion**.
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** false positive
- **Diagnostic reviewer root cause:** pre-Doctor equivalence: numeric match: model has \$408,419 ~ GT 408,419
- **Verifier verdict:** false positive (confidence: high)
- **Validator flags:** Unsupported number: '\$408.4' does not appear in the evidence

Failure 70 — `SEC\_10K\_multi\_1265\_01` · HIGH confidence

BAD QUESTION

- **Category:** Cross-Reference
- **Question:** What is the net liquidation loss for securitized loans in 2025, and how does it compare to the net liquidation loss in 2024?
- ✓ **Ground truth:** The net liquidation loss for securitized loans in 2025 is \$5,233 million, which is higher than the net liquidation loss in 2024 of \$2,329 million.
- ✗ **Model answered:** The net liquidation loss for securitized loans was **\$301 million in 2025**, compared with **\$89 million in 2024**-an increase of **\$212 million**.
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** bad test

- **Diagnostic reviewer root cause:** The ground truth read the adjacent '90 days past due' columns (\$5,233 million and \$2,329 million) instead of the 'Net liquidation losses/(recoveries)' columns. The model correctly used \$301 million and \$89 million and derived the \$212 million increase; the validator incorrectly treated this arithmetic result as unsupported because it does not appear verbatim.
- **Verifier verdict:** bad test (confidence: high)
- **Validator flags:** Unsupported number: '\$212' does not appear in the evidence

Failure 71 — `SEC\_10K\_multi\_0231\_00` · — confidence

DETERMINISTIC:GT NUMBER CONFLICT

- **Severity:** HIGH — an incorrect quantitative result
- **Category:** Precision
- **Question:** What is the total credit provided and capital raised for U.S. small businesses?
- ✓ **Ground truth:** \$2.9 trillion
- ✗ **Model answered:** The total credit provided and capital raised for U.S. small businesses was **\$33 billion**.
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** model error
- **Diagnostic reviewer root cause:** [code:gt_number_conflict] Short numeric GT; response asserts a different quantity
- **Verifier verdict:** — (confidence: —)
- **Validator flags:** (none)

Failure 72 — `SEC\_10K\_multi\_0574\_01` · HIGH confidence

VALIDATOR OVER-FIRED

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Precision
- **Question:** What is the total number of government guaranteed loans at December 31, 2025?
- ✓ **Ground truth:** 1,159
- ✗ **Model answered:** The total government-guaranteed loans at December 31, 2025 were **\$1.159 billion**.
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** false positive
- **Diagnostic reviewer root cause:** Validator failed to recognize the unit conversion: the table reports \$1,159 million, which is equivalent to \$1.159 billion.
- **Verifier verdict:** false positive (confidence: high)
- **Validator flags:** Unsupported number: '\$1.159' does not appear in the evidence

Failure 73 — `SEC\_10K\_multi\_0911\_00` · HIGH confidence

UNIT/FORMAT MISMATCH

- **Category:** Cross-Reference
- **Question:** What is the total amount for Insurance and what is the total amount for Central Govt, and how do these two totals compare?
- ✓ **Ground truth:** The total amount for Insurance is \$25,031 and the total amount for Central Govt is \$15,164. These two totals are not comparable as they are different categories.
- ✗ **Model answered:** As of December 31, 2025:
 - **Insurance:** \$25.031 billion
 - **Central Govt:** \$15.164 billion

Insurance exposure was **\$9.867 billion higher**, approximately **65% more** than Central Govt exposure.

- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** bad test

- **Diagnostic reviewer root cause:** The model converted the table figures from millions to billions and then performed the numerical comparison requested; the validator could not recognize the scaled-unit equivalence. However, the model's claim that Insurance is 65% higher conflicts with the ground truth's claim that the categories are not comparable.
- **Verifier verdict:** bad test (confidence: high)
- **Validator flags:** Unsupported number: '\$15.164' does not appear in the evidence

Failure 74 — `SEC\_10K\_multi\_neg\_0442\_01` · — confidence

DETERMINISTIC:ANSWERED UNANSWERABLE

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What is the total amount of regulatory capital that must be held by the Firm's principal IDI subsidiary, JPMorgan Chase Bank, N.A.?
- **Ground truth:** UNANSWERABLE
- **Model answered:** JPMorgan Chase Bank, N.A.'s total capital requirement was **10.5% of risk-weighted assets**. Using its 2025 Standardized RWA of **\$1.928 trillion**, this equates to approximately **\$202.4 billion** in required regulatory capital. It actually held **\$317.7 billion** of Total capital under the Standardized approach.
- **Diagnostic reviewer diagnosis:** **Negative-answer specialist** model error
- **Diagnostic reviewer root cause:** [code:answered_unanswerable] GT is UNANSWERABLE but the response asserts a concrete answer
- **Verifier verdict:** — (confidence: —)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer; Unsupported number: '\$1.928' does not appear in the evidence

Failure 75 — `SEC\_10K\_multi\_neg\_0646\_04` · — confidence

DETERMINISTIC:ANSWERED UNANSWERABLE

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What is the average allowance for loan losses to retained nonaccrual loans?
- **Ground truth:** UNANSWERABLE
- **Model answered:** The average allowance for loan losses to retained nonaccrual loans was **325%**, based on **311% in 2025** and **339% in 2024**.
- **Diagnostic reviewer diagnosis:** **Negative-answer specialist** model error
- **Diagnostic reviewer root cause:** [code:answered_unanswerable] GT is UNANSWERABLE but the response asserts a concrete answer
- **Verifier verdict:** — (confidence: —)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer; Unsupported number: '325%' does not appear in the evidence

Failure 76 — `SEC\_10K\_multi\_0642\_00` · HIGH confidence

VALIDATOR OVER-FIRED

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Cross-Reference
- **Question:** What is the net charge-off for the consumer credit card segment in the year ended December 31, 2024?

- ✓ **Ground truth:** 7,142
- ✗ **Model answered:** The consumer credit card segment's net charge-offs were **\$7.142 billion** for the year ended December 31, 2024.
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** false positive
- **Diagnostic reviewer root cause:** Validator failed to recognize unit conversion: the table reports 7,142 in millions, which is equivalent to \$7.142 billion.
- **Verifier verdict:** false positive (confidence: high)
- **Validator flags:** Unsupported number: '\$7.142' does not appear in the evidence

Failure 77 — `SEC\_10K\_multi\_0803\_01` · HIGH confidence

MISINTERPRETED

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Precision
- **Question:** What is the total number of debt instruments?
- ✓ **Ground truth:** 2,138
- ✗ **Model answered:** Total debt instruments were **\$2,182 million (\$2.182 billion)** as of December 31, 2025.
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** model error
- **Diagnostic reviewer root cause:** The model misread or transposed the table's total, reporting 2,182 instead of the stated 2,138, and added an unsupported date context.
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Unsupported number: '\$2.182' does not appear in the evidence

Failure 78 — `SEC\_10K\_multi\_1157\_00` · HIGH confidence

TABLE MISREAD

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Precision
- **Question:** What is the total retained loans for 2022?
- ✓ **Ground truth:** \$14,917
- ✗ **Model answered:** The total retained loans for the 2022 origination year were **\$5.069 billion**.
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** model error
- **Diagnostic reviewer root cause:** Model read or derived a value from the wrong table entry/source instead of using the 2022 column in the 'Total retained loans' row, which shows \$14,917 million.
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Unsupported number: '\$5.069' does not appear in the evidence

Failure 79 — `SEC\_10K\_multi\_0751\_01` · HIGH confidence

TABLE MISREAD

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Precision
- **Question:** What is the total U.S. Treasury and government agencies?
- ✓ **Ground truth:** 234,779
- ✗ **Model answered:** The total U.S. Treasury and government agencies available-for-sale securities was **\$315.822 billion in fair value** as of December 31, 2025 (**\$313.470 billion amortized cost**).
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** model error

- **Diagnostic reviewer root cause:** The model pulled figures from the wrong table or reporting period and mischaracterized them as available-for-sale fair value and amortized cost, instead of summing the evidence row's components ($234,491 + 288 = 234,779$).
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Unsupported number: '\$315.822' does not appear in the evidence

Failure 80 — `SEC\_10K\_multi\_neg\_0744\_03` · — confidence

DETERMINISTIC:ANSWERED UNANSWERABLE

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What is the total amount of accounts payable and other liabilities measured at fair value on a recurring basis?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** The total was **\$5.893 billion (\$5,893 million)**.
- **Diagnostic reviewer diagnosis:** **Negative-answer specialist** model error
- **Diagnostic reviewer root cause:** [code:answered_unanswerable] GT is UNANSWERABLE but the response asserts a concrete answer
- **Verifier verdict:** — (confidence: —)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer; Unsupported number: '\$5.893' does not appear in the evidence

Failure 81 — `SEC\_10K\_multi\_neg\_0539\_02` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** How many credit rating downgrades occurred for JPMorgan Chase & Co. in 2025?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** Zero credit rating downgrades occurred for JPMorgan Chase & Co. in 2025.
- **Diagnostic reviewer diagnosis:** **Negative-answer specialist** model error
- **Diagnostic reviewer root cause:** The model inferred that no downgrades occurred from a passage that only discusses downgrade risks and lists ratings as of December 31, 2025; a point-in-time ratings table does not establish the number of downgrades during the year.
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 82 — `SEC\_10K\_multi\_neg\_0109\_06` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What is the name of the chairman of the board of directors of JPMorgan Chase & Co.?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** James Dimon.
- **Diagnostic reviewer diagnosis:** **Negative-answer specialist** model error
- **Diagnostic reviewer root cause:** The model relied on external knowledge about JPMorgan Chase's leadership rather than restricting its answer to the provided evidence, which does not identify the chairman.

- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 83 — `SEC\_10K\_multi\_0082\_01` · HIGH confidence

VALIDATOR OVER-FIRED

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Precision
- **Question:** What is the amount of cash paid for income taxes (net of refunds) for the fiscal year ended January 25, 2026?
- ✓ **Ground truth:** \$20,288
- ✗ **Model answered:** The amount of cash paid for income taxes, net of refunds, was **\$20.288 billion** for the fiscal year ended January 25, 2026.
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** false positive
- **Diagnostic reviewer root cause:** Validator failed to recognize a valid unit conversion: the table reports \$20,288 million, which equals \$20.288 billion.
- **Verifier verdict:** false positive (confidence: high)
- **Validator flags:** Unsupported number: '\$20.288' does not appear in the evidence

Failure 84 — `SEC\_10K\_multi\_0109\_00` · — confidence

DETERMINISTIC: DIRECTION CONFLICT

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Conditional
- **Question:** If JPMorganChase is a large accelerated filer, does it have to file reports pursuant to Section 13 or 15(d) of the Act?
- ✓ **Ground truth:** No
- ✗ **Model answered:** Yes. JPMorganChase is a large accelerated filer and indicated that it is required to file reports pursuant to Section 13 or 15(d) of the Act.
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** model error
- **Diagnostic reviewer root cause:** [code:direction_conflict] GT is 'no' but response leads 'yes'
- **Verifier verdict:** — (confidence: —)
- **Validator flags:** Direction mismatch: GT negative, response affirmative

Failure 85 — `SEC\_10K\_multi\_1160\_00` · HIGH confidence

MISINTERPRETED

- **Category:** Cross-Reference
- **Question:** What is the total retained loans amount in the U.S. and the total retained loans amount in non-U.S., respectively?
- ✓ **Ground truth:** \$165,545 and \$3,167
- ✗ **Model answered:** As of December 31, 2025, total retained loans were **\$626.060 billion in the U.S.** and **\$166.307 billion in non-U.S.** markets, respectively.
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** bad test
- **Diagnostic reviewer root cause:** [verifier] The ground truth is incorrect, as it uses the value for "Total retained loans" (\$165,545) as the answer for "Total U.S.", when the evidence clearly provides a different value for "Total U.S." (\$162,378).
- **Verifier verdict:** bad test (confidence: high)
- **Validator flags:** Unsupported number: '\$626.060' does not appear in the evidence

Failure 86 — `SEC\_10K\_multi\_neg\_0034\_05` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What is the average maturity date of the notes issued by Apple Inc. in 2025?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** The notes mature in 2028, 2030, 2032, and 2035. Their simple average maturity year is **2031.25**, or approximately **2031**.
- **Diagnostic reviewer diagnosis:** **Negative-answer specialist** model error
- **Diagnostic reviewer root cause:** [verifier] The model answered a question that is unanswerable from the evidence by assuming a simple average was the correct calculation method, when the evidence lacks the principal amounts of each note required for a standard weighted average maturity calculation.
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 87 — `SEC\_10K\_multi\_1297\_01` · HIGH confidence

VALIDATOR OVER-FIRED

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Cross-Reference
- **Question:** What is the net rental expense for the year ended December 31, 2025, and how does it compare to the gross rental expense?
- ✓ **Ground truth:** Net rental expense: \$2,360, Gross rental expense: \$2,388
- ✗ **Model answered:** For the year ended December 31, 2025, net rental expense was **\$2.360 billion**, compared with gross rental expense of **\$2.388 billion**. The **\$28 million difference** reflects sublease rental income.
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** false positive
- **Diagnostic reviewer root cause:** Validator failed to recognize the unit conversion from \$2,360 million in the evidence to \$2.360 billion in the response; the gross expense and \$28 million sublease-income difference also match.
- **Verifier verdict:** false positive (confidence: high)
- **Validator flags:** Unsupported number: '\$2.360' does not appear in the evidence

Failure 88 — `SEC\_10K\_multi\_0388\_00` · HIGH confidence

UNSUPPORTED NUMBER

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Precision
- **Question:** What is the allowance for lending-related commitments?
- ✓ **Ground truth:** 43
- ✗ **Model answered:** At December 31, 2025, the allowance for lending-related commitments was **\$5.071 billion**; **\$83 million** for consumer excluding credit card, **\$2.2 billion** for credit card, and **\$2.788 billion** for wholesale.
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** model error
- **Diagnostic reviewer root cause:** Model used an unrelated table or disclosure and summed consumer, credit-card, and wholesale figures instead of reading the 'Allowance for lending-related commitments' row, whose relevant value is 43.

- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Unsupported number: '\$5.071' does not appear in the evidence

Failure 89 — `SEC\_10K\_multi\_neg\_0254\_04` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What is the name of the specific non-GAAP financial measure used by the Firm to compare with competitors?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** The non-GAAP financial measure is **managed basis**.
- **Diagnostic reviewer diagnosis:** **Negative-answer specialist** model error
- **Diagnostic reviewer root cause:** The model conflated the named “managed basis” measure, which the passage says supports year-to-year revenue comparability, with the unnamed “other non-GAAP financial measures” used to facilitate competitor comparisons.
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 90 — `SEC\_10K\_multi\_0055\_01` · HIGH confidence

BAD QUESTION

- **Category:** Cross-Reference
- **Question:** What is the percentage increase in revenue for fiscal year 2026 compared to the previous year, and how does it relate to the growth in data center revenue?
- ✓ **Ground truth:** 65% increase in revenue, which is higher than the 68% growth in data center revenue.
- ✗ **Model answered:** Fiscal year 2026 revenue increased **65%** to **\$215.9 billion** from \$130.5 billion. Data Center revenue grew slightly faster, rising **68% year over year**, driven by accelerated computing, AI, and strong Blackwell demand.
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** bad test
- **Diagnostic reviewer root cause:** The model correctly compared the stated growth rates and concluded that Data Center revenue grew faster (68% versus 65%), whereas the provided correct answer reverses that relationship. The validator separately failed to recognize that \$130.5 billion is the rounded equivalent of the evidence's \$130,497 million.
- **Verifier verdict:** bad test (confidence: high)
- **Validator flags:** Unsupported number: '\$130.5' does not appear in the evidence

Failure 91 — `SEC\_10K\_multi\_0929\_01` · HIGH confidence

BAD QUESTION

- **Category:** Calculation
- **Question:** If the gross long notional derivative contracts are \$500 and the gross short notional derivative contracts are \$300, what is the total?
- ✓ **Ground truth:** 200
- ✗ **Model answered:** 800
- **Diagnostic reviewer diagnosis:** **Calculation specialist** bad test
- **Diagnostic reviewer root cause:** The model correctly interpreted “total” and the evidence’s definition as the sum of gross long and gross short notionals: $500 + 300 = 800$. The ground truth instead appears to have subtracted the short amount from the long amount to get 200.
- **Verifier verdict:** bad test (confidence: high)

- **Validator flags:** Number mismatch: expected 200, not found in response

Failure 92 — `SEC\_10K\_multi\_0551\_01` · HIGH confidence

TABLE MISREAD

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Precision
- **Question:** What was the total nonaccrual loans at December 31, 2024?
- ✓ **Ground truth:** 8,837
- ✗ **Model answered:** Total nonaccrual loans at December 31, 2024 were **\$4.760 billion**.
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** model error
- **Diagnostic reviewer root cause:** Model extracted or computed a figure from the wrong table or loan category instead of reading the Firmwide total nonaccrual loans row for 2024, which states \$8,837 million.
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Unsupported number: '\$4.760' does not appear in the evidence

Failure 93 — `SEC\_10K\_multi\_1497\_01` · HIGH confidence

VALIDATOR OVER-FIRED

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Cross-Reference
- **Question:** What was the average interest rate earned on securities borrowed in U.S. operations for the year ended December 31, 2025, and what was the average balance of securities borrowed in U.S. operations for the same year?
- ✓ **Ground truth:** 3.92% and \$185,559
- ✗ **Model answered:** The average interest rate was **3.92%**, and the average balance was **\$185.559 billion** (or **\$185,559 million**).
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** false positive
- **Diagnostic reviewer root cause:** pre-Doctor equivalence: numeric match: model has 3.92% ~ GT 3.92% and \$185,559
- **Verifier verdict:** false positive (confidence: high)
- **Validator flags:** Unsupported number: '\$185.559' does not appear in the evidence

Failure 94 — `SEC\_10K\_multi\_1501\_00` · HIGH confidence

UNSUPPORTED NUMBER

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Precision
- **Question:** What is the amount of noninterest-bearing liabilities funding interest-earning assets?
- ✓ **Ground truth:** 284,913
- ✗ **Model answered:** The amount of noninterest-bearing liabilities funding interest-earning assets was **\$670.426 billion** in 2025.
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** model error
- **Diagnostic reviewer root cause:** The model selected an unrelated 2025 liability balance, likely from a different table row or section, instead of the footnote-linked average balance of 284,913 million for noninterest-bearing liabilities funding interest-earning assets.
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Unsupported number: '\$670.426' does not appear in the evidence

Failure 95 — `SEC\_10K\_multi\_0487\_01` · — confidence

DETERMINISTIC:GT NUMBER CONFLICT

- **Severity:** HIGH — an incorrect quantitative result
- **Category:** Precision
- **Question:** What is the amount of the allowance for credit losses in 2025?
- ✓ **Ground truth:** 468 million
- ✗ **Model answered:** The allowance for credit losses was **\$31.2 billion as of December 31, 2025**.
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** model error
- **Diagnostic reviewer root cause:** [code:gt_number_conflict] Short numeric GT; response asserts a different quantity
- **Verifier verdict:** — (confidence: —)
- **Validator flags:** (none)

Failure 96 — `SEC\_10K\_multi\_neg\_1537\_00` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What is the maximum number of employees covered by collective bargaining agreements at our key manufacturing locations?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** Approximately 400 employees in North America were covered by collective bargaining agreements as of December 31, 2025.
- **Diagnostic reviewer diagnosis:** **Negative-answer specialist** model error
- **Diagnostic reviewer root cause:** The model hallucinated a specific employee count and region/date, likely importing details from another filing or context; the evidence only says that many employees are covered and provides no number or maximum.
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 97 — `SEC\_10K\_multi\_neg\_0052\_04` · — confidence

DETERMINISTIC:ANSWERED UNANSWERABLE

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** How many lawsuits are pending against the company at any given time?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** Approximately 15,000 active claims are pending against Aqua-Chem.
- **Diagnostic reviewer diagnosis:** **Negative-answer specialist** model error
- **Diagnostic reviewer root cause:** [code:answered_unanswerable] GT is UNANSWERABLE but the response asserts a concrete answer
- **Verifier verdict:** — (confidence: —)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 98 — `SEC\_10K\_multi\_0060\_00` · HIGH confidence

VALIDATOR OVER-FIRED

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Precision
- **Question:** What is the amount of cash and cash equivalents as of January 25, 2026?
- ✓ **Ground truth:** 10,605
- ✗ **Model answered:** Cash and cash equivalents were **\$10.605 billion** as of January 25, 2026.
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** false positive
- **Diagnostic reviewer root cause:** Validator failed to account for unit conversion: the table reports \$10,605 million, which is exactly \$10.605 billion as stated by the model.
- **Verifier verdict:** false positive (confidence: high)
- **Validator flags:** Unsupported number: '\$10.605' does not appear in the evidence

Failure 99 — `SEC\_10K\_multi\_1025\_01` · HIGH confidence

UNSUPPORTED NUMBER

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Precision
- **Question:** What is the total notional amount?
- ✓ **Ground truth:** 628,130
- ✗ **Model answered:** The total derivative notional amount was **\$50.642 trillion (\$50,642 billion)** as of December 31, 2025.
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** model error
- **Diagnostic reviewer root cause:** Model used an unrelated derivative figure from the wrong reporting period/table, citing December 31, 2025 instead of extracting the evidence's total notional amount of 628,130 for December 31, 2024.
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Unsupported number: '\$50.642' does not appear in the evidence

Failure 100 — `SEC\_10K\_multi\_0520\_01` · HIGH confidence

UNIT/FORMAT MISMATCH

- **Category:** Precision
- **Question:** What is the total short-term unsecured funding for 2025?
- ✓ **Ground truth:** 28,668
- ✗ **Model answered:** The total short-term unsecured funding for 2025 was **\$27.341 billion**.
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** bad test
- **Diagnostic reviewer root cause:** The model selected the December 31, 2025 point-in-time balance (\$27,341 million) rather than the 2025 average balance (\$28,668 million); its billion-unit conversion was valid and supported by the table.
- **Verifier verdict:** bad test (confidence: high)
- **Validator flags:** Unsupported number: '\$27.341' does not appear in the evidence

Failure 101 — `SEC\_10K\_multi\_0781\_01` · HIGH confidence

TABLE MISREAD

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Cross-Reference
- **Question:** What is the total value of debt instruments and equity securities combined?
- ✓ **Ground truth:** $2,144 + 62 = 2,206$
- ✗ **Model answered:** The combined value was **\$549.156 billion** (\$417.428 billion in debt instruments plus \$131.728 billion in equity securities).
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** model error
- **Diagnostic reviewer root cause:** Model pulled debt and equity values from the wrong table or unrelated context instead of adding the evidence figures 2,144 and 62, producing an unsupported \$549.156 billion total.
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Unsupported number: '\$549.156' does not appear in the evidence

Failure 102 — `SEC\_10K\_multi\_0597\_00` · HIGH confidence

BAD QUESTION

- **Category:** Precision
- **Question:** What is the threshold for Chemicals & Plastics?
- ✓ **Ground truth:** 8,152
- ✗ **Model answered:** No specific threshold is stated for Chemicals & Plastics. Its 2025 wholesale credit exposure was **\$23.79 billion**, including **\$2.091 billion criticized performing** and **\$93 million criticized nonperforming**.
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** bad test
- **Diagnostic reviewer root cause:** The model used unrelated wholesale credit-exposure figures, likely from a different table or reporting period, instead of extracting 8,152 from the Chemicals & Plastics row's threshold column.
- **Verifier verdict:** bad test (confidence: high)
- **Validator flags:** Unsupported number: '\$2.091' does not appear in the evidence

Failure 103 — `SEC\_10K\_multi\_neg\_0254\_01` · — confidence

DETERMINISTIC:ANSWERED UNANSWERABLE

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What is the average annual revenue growth rate on a managed basis?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** Managed-basis total revenue was:
 - **2023:** \$162.366 billion
 - **2025:** \$185.581 billion

Average annual growth rate (CAGR), 2023-2025:

$$\left[\left(\frac{185.581}{162.366} \right)^{\frac{1}{2}} - 1 \right] \approx \mathbf{6.9\%}$$

- **Diagnostic reviewer diagnosis:** **Negative-answer specialist** model error
- **Diagnostic reviewer root cause:** [code:answered_unanswerable] GT is UNANSWERABLE but the response asserts a concrete answer
- **Verifier verdict:** — (confidence: —)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer; Unsupported number: '\$162.366' does not appear in the evidence

Failure 104 — `SEC\_10K\_multi\_neg\_1554\_04` · — confidence

DETERMINISTIC:ANSWERED UNANSWERABLE

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What was the percentage of the net operating revenue increase attributed to acquisitions and divestitures in the Asia Pacific operating segment?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** Acquisitions and divestitures had a **1% unfavorable impact** on Asia Pacific net operating revenues in 2025 versus 2024.
- **Diagnostic reviewer diagnosis:** **Negative-answer specialist** model error
- **Diagnostic reviewer root cause:** [code:answered_unanswerable] GT is UNANSWERABLE but the response asserts a concrete answer
- **Verifier verdict:** — (confidence: —)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 105 — `SEC\_10K\_multi\_neg\_0727\_01` · — confidence

DETERMINISTIC:ANSWERED UNANSWERABLE

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What is the specific percentage of loans carried at fair value that are classified as level 2?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** Approximately **95.7%** of loans carried at fair value are classified as Level 2:

$$\left[\frac{\$70.684 \text{B} - \$3.062 \text{B}}{\$70.684 \text{B}} \times 100 \approx 95.7\% \right]$$

- **Diagnostic reviewer diagnosis:** **Negative-answer specialist** model error
- **Diagnostic reviewer root cause:** [code:answered_unanswerable] GT is UNANSWERABLE but the response asserts a concrete answer
- **Verifier verdict:** — (confidence: —)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer; Unsupported number: '95.7%' does not appear in the evidence

Failure 106 — `SEC\_10K\_multi\_neg\_1614\_01` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What is the date of the next annual report submission?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** The next annual submission is due **April 6, 2026**.
- **Diagnostic reviewer diagnosis:** **Negative-answer specialist** model error
- **Diagnostic reviewer root cause:** The model inferred or fabricated a future filing deadline from the report's February 20, 2026 signature date, even though the evidence provides no next annual report submission date or sufficient filing-deadline information.
- **Verifier verdict:** model error (confidence: high)

- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 107 — `SEC\_10K\_multi\_0503\_01` · HIGH confidence

VALIDATOR OVER-FIRED

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Cross-Reference
- **Question:** What is the net capital of J.P. Morgan Securities as of December 31, 2025, and how does it compare to the minimum requirements?
- ✓ **Ground truth:** \$27,196 million, which is greater than the minimum of \$6,559 million.
- ✗ **Model answered:** As of December 31, 2025, J.P. Morgan Securities had **net capital of \$27.196 billion**, compared with a **minimum requirement of \$6.559 billion**-an excess of **\$20.637 billion**.
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** false positive
- **Diagnostic reviewer root cause:** Validator failed to recognize a unit conversion: the evidence reports \$6,559 million, which is exactly \$6.559 billion. The model also correctly converted \$27,196 million to \$27.196 billion and calculated the \$20.637 billion excess.
- **Verifier verdict:** false positive (confidence: high)
- **Validator flags:** Unsupported number: '\$6.559' does not appear in the evidence

Failure 108 — `SEC\_10K\_multi\_1588\_01` · HIGH confidence

VALIDATOR OVER-FIRED

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Cross-Reference
- **Question:** What is the total revenue from concentrate operations in the United States for the year ended December 31, 2025, and how does it compare to the total revenue from finished product operations in the same period?
- ✓ **Ground truth:** The total revenue from concentrate operations in the United States for the year ended December 31, 2025, is \$8,956 million, and the total revenue from finished product operations is \$10,171 million.
- ✗ **Model answered:** For 2025 in the United States:
 - **Concentrate operations revenue:** \$8,956 million
 - **Finished product operations revenue:** \$10,171 million

Finished product operations generated **\$1,215 million more** revenue than concentrate operations.

- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** false positive
- **Diagnostic reviewer root cause:** The model correctly quoted both revenue figures and added a valid comparison by subtracting \$8,956 million from \$10,171 million. The validator only checked for verbatim numeric support and could not recognize the derived \$1,215 million difference.
- **Verifier verdict:** false positive (confidence: high)
- **Validator flags:** Unsupported number: '\$1,215' does not appear in the evidence

Failure 109 — `SEC\_10K\_multi\_neg\_1554\_00` · — confidence

DETERMINISTIC:ANSWERED UNANSWERABLE

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What was the total concentrate sales volume growth rate for the Asia Pacific operating segment?

- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** The Asia Pacific operating segment's concentrate sales volume was **even (0%)** in 2025 versus 2024. After structural changes, it grew **1%**.
- **Diagnostic reviewer diagnosis:** **Negative-answer specialist** model error
- **Diagnostic reviewer root cause:** [code:answered_unanswerable] GT is UNANSWERABLE but the response asserts a concrete answer
- **Verifier verdict:** — (confidence: —)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer; Unsupported number: '0%' does not appear in the evidence

Failure 110 — `SEC\_10K\_multi\_1566\_01` · HIGH confidence

VALIDATOR OVER-FIRED

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Cross-Reference
- **Question:** How did the operating margin in 2025 compare to the operating margin in 2024 for the EMEA operating segment?
- ✓ **Ground truth:** The operating margin in 2025 was 39.7%, while in 2024 it was 41.4%. This indicates a decrease in operating margin for the EMEA operating segment from 2024 to 2025.
- ✗ **Model answered:** The EMEA operating margin decreased from 41.4% in 2024 to 39.7% in 2025, a decline of 1.7 percentage points.
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** false positive
- **Diagnostic reviewer root cause:** The model correctly quoted both margins and computed their difference (41.4% - 39.7% = 1.7 percentage points). The validator only checked whether the derived number appeared verbatim in the evidence and failed to recognize valid arithmetic.
- **Verifier verdict:** false positive (confidence: high)
- **Validator flags:** Unsupported number: '1.7' does not appear in the evidence

Failure 111 — `SEC\_10K\_multi\_1286\_00` · HIGH confidence

VALIDATOR OVER-FIRED

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Cross-Reference
- **Question:** What is the total amortization expense for the years ended December 31, 2025 and 2024, and how does it relate to the net carrying values of other intangible assets?
- ✓ **Ground truth:** The total amortization expense was \$292 million for 2025 and \$339 million for 2024, which relates to the net carrying values of \$1.3 billion and \$1.7 billion, respectively.
- ✗ **Model answered:** Amortization expense totaled **\$631 million: \$292 million in 2025 plus \$339 million in 2024.**

This expense applies to finite-lived intangible assets, whose net carrying values were **\$1.3 billion in 2025** and **\$1.7 billion in 2024**. Indefinite-lived intangible assets-**\$1.3 billion and \$1.2 billion**, respectively-are not amortized. Total net other intangible assets were therefore **\$2.6 billion in 2025** and **\$2.9 billion in 2024**.

- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** false positive
- **Diagnostic reviewer root cause:** The model correctly gave the 2025 and 2024 amortization expenses and related finite-lived net carrying values, but it also volunteered total net intangible values derived by summing the rounded finite- and indefinite-lived amounts. The validator only checked for numbers explicitly appearing in the evidence and could not recognize the derived \$2.6 billion figure.
- **Verifier verdict:** false positive (confidence: high)
- **Validator flags:** Unsupported number: '\$2.6' does not appear in the evidence

Failure 112 — `SEC\_10K\_multi\_1156\_01` · HIGH confidence

BAD QUESTION

- **Category:** Precision
- **Question:** What is the total retained loans for noninvestment-grade risk ratings as of December 31, 2024?
- ✓ **Ground truth:** 162,432
- ✗ **Model answered:** As of December 31, 2024, total retained loans rated noninvestment-grade were **\$218.726 billion**.
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** bad test
- **Diagnostic reviewer root cause:** The model appears to have aggregated noninvestment-grade retained loans across multiple loan categories rather than selecting the secured-by-real-estate table's "Total retained loans" row used by the ground truth.
- **Verifier verdict:** bad test (confidence: high)
- **Validator flags:** Unsupported number: '\$218.726' does not appear in the evidence

Failure 113 — `SEC\_10K\_multi\_neg\_1274\_02` · — confidence

DETERMINISTIC:ANSWERED UNANSWERABLE

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What is the total value of servicer advances at December 31?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** Servicer advances totaled **\$493 million at December 31, 2025**, net of an allowance for uncollectible amounts.
- **Diagnostic reviewer diagnosis:** **Negative-answer specialist** model error
- **Diagnostic reviewer root cause:** [code:answered_unanswerable] GT is UNANSWERABLE but the response asserts a concrete answer
- **Verifier verdict:** — (confidence: —)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 114 — `SEC\_10K\_multi\_0583\_01` · HIGH confidence

BAD QUESTION

- **Category:** Cross-Reference
- **Question:** What is the total amount of derivative receivables, net of collateral, and what percentage of total derivative receivables is that?
- ✓ **Ground truth:** 11,515, 75%
- ✗ **Model answered:** As of December 31, 2025, derivative receivables net of all collateral were **\$27.937 billion**, equal to approximately **48.4%** of total net derivative receivables of \$57.777 billion.
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** bad test
- **Diagnostic reviewer root cause:** Model used figures from the wrong reporting period or source (December 2025 rather than the December 2024 table) and computed an unrelated percentage.
- **Verifier verdict:** bad test (confidence: high)
- **Validator flags:** Unsupported number: '\$27.937' does not appear in the evidence

Failure 115 — `SEC\_10K\_multi\_1255\_00` · HIGH confidence

UNIT/FORMAT MISMATCH

- **Category:** Precision

- **Question:** What is the total proceeds received from loan sales in 2023?
- ✓ **Ground truth:** 19,389
- ✗ **Model answered:** The total proceeds received from loan sales in 2023 were **\$19.689 billion**.
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** bad test
- **Diagnostic reviewer root cause:** The model calculated total proceeds by adding 2023 cash proceeds of \$300 million to securities proceeds of \$19,389 million, yielding \$19,689 million (\$19.689 billion), while the ground truth incorrectly reports only the securities component.
- **Verifier verdict:** bad test (confidence: high)
- **Validator flags:** Unsupported number: '\$19.689' does not appear in the evidence

Failure 116 — `SEC\_10K\_multi\_neg\_0080\_01` · — confidence

DETERMINISTIC:ANSWERED UNANSWERABLE

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What is the estimated range of loss for the legal proceedings described in the passage?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** The estimated aggregate range of reasonably possible losses is **\$0 to approximately \$1.2 billion**, in excess of established reserves, as of December 31, 2025.
- **Diagnostic reviewer diagnosis:** **Negative-answer specialist** model error
- **Diagnostic reviewer root cause:** [code:answered_unanswerable] GT is UNANSWERABLE but the response asserts a concrete answer
- **Verifier verdict:** — (confidence: —)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 117 — `SEC\_10K\_multi\_neg\_0003\_01` · HIGH confidence

VALIDATOR OVER-FIRED

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What is the average cost of an industrial accident at one of Apple's suppliers?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** Apple does not disclose an average cost. It states that an industrial accident at a supplier could cause injuries or loss of life, business disruption, reputational harm, recovery expenditures, and lost sales.
- **Diagnostic reviewer diagnosis:** **Negative-answer specialist** false positive
- **Diagnostic reviewer root cause:** The model correctly recognized that the filing provides no average accident cost and expressed the unanswerability in prose; the validator misclassified the added summary of possible consequences as an attempted answer.
- **Verifier verdict:** false positive (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 118 — `SEC\_10K\_multi\_1409\_00` · — confidence

DETERMINISTIC:GT NUMBER CONFLICT

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Cross-Reference

- **Question:** What was the total amount for standby letters of credit and other financial guarantees as of December 31, 2024?
- ✓ **Ground truth:** \$9.5 billion
- ✗ **Model answered:** The total contractual amount was **\$28.676 billion** as of December 31, 2024.
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** model error
- **Diagnostic reviewer root cause:** [code:gt_number_conflict] Short numeric GT; response asserts a different quantity
- **Verifier verdict:** — (confidence: —)
- **Validator flags:** Unsupported number: '\$28.676' does not appear in the evidence

Failure 119 — `SEC\_10K\_multi\_0932\_00` · HIGH confidence

UNSUPPORTED NUMBER

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Precision
- **Question:** What is the total fair value of trading assets and liabilities?
- ✓ **Ground truth:** 606,382
- ✗ **Model answered:** As of December 31, 2025, the total fair value was:
 - **Trading assets:** \$606.382 billion gross; **\$57.777 billion net**
 - **Trading liabilities:** \$596.738 billion gross; **\$46.329 billion net**
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** model error
- **Diagnostic reviewer root cause:** [verifier] The model's answer contains a factual error, stating the date is "December 31, 2025" when the evidence clearly shows the date is "December 31, 2024".
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Unsupported number: '\$606.382' does not appear in the evidence

Failure 120 — `SEC\_10K\_multi\_1025\_00` · HIGH confidence

BAD QUESTION

- **Category:** Precision
- **Question:** What is the fair value of receivables?
- ✓ **Ground truth:** 190,662
- ✗ **Model answered:** As of December 31, 2025, derivative receivables had a **gross fair value of \$606.382 billion** and a **net fair value of \$57.777 billion** after netting adjustments.
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** bad test
- **Diagnostic reviewer root cause:** The model used figures from a different derivatives table or disclosure and reported 2025 gross and net derivative receivables rather than the requested table value.
- **Verifier verdict:** bad test (confidence: high)
- **Validator flags:** Unsupported number: '\$606.382' does not appear in the evidence

Failure 121 — `SEC\_10K\_multi\_1029\_00` · HIGH confidence

VALIDATOR OVER-FIRED

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Precision
- **Question:** What is the total underwriting fee for 2023?
- ✓ **Ground truth:** 3,759
- ✗ **Model answered:** The total underwriting fee for 2023 was **\$3.759 billion**.

- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** false positive
- **Diagnostic reviewer root cause:** Validator failed to recognize a valid unit conversion: the table reports \$3,759 million, which equals \$3.759 billion.
- **Verifier verdict:** false positive (confidence: high)
- **Validator flags:** Unsupported number: '\$3.759' does not appear in the evidence

Failure 122 — `SEC\_10K\_multi\_0630\_00` · HIGH confidence

UNSUPPORTED NUMBER

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Precision
- **Question:** What is the total for wholesale loans?
- **Ground truth:** 843,374
- **Model answered:** Total wholesale loans were **\$843.374 billion** as of December 31, 2025.
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** model error
- **Diagnostic reviewer root cause:** [verifier] The model's answer fabricated the date "as of December 31, 2025," which is not mentioned anywhere in the provided evidence.
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Unsupported number: '\$843.374' does not appear in the evidence

Failure 123 — `SEC\_10K\_multi\_neg\_0109\_03` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What is the name of the CEO of JPMorgan Chase & Co.?
- **Ground truth:** UNANSWERABLE
- **Model answered:** James Dimon.
- **Diagnostic reviewer diagnosis:** **Negative-answer specialist** model error
- **Diagnostic reviewer root cause:** The model relied on external/general knowledge that JPMorgan Chase's CEO is James Dimon instead of restricting its answer to the supplied evidence, which does not identify the CEO.
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 124 — `SEC\_10K\_multi\_1584\_00` · HIGH confidence

UNIT/FORMAT MISMATCH

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Cross-Reference
- **Question:** What is the total amount of cash and cash equivalents at the end of 2025, considering the cash and cash equivalents at the beginning of the year and the net increase (decrease) in cash, cash equivalents, restricted cash and restricted cash equivalents during the year?
- **Ground truth:** 10,270
- **Model answered:** Beginning balance of \$11.488 billion minus the \$478 million decrease equals **\$11.010 billion** in total cash, cash equivalents, and restricted cash. After subtracting \$740 million of restricted cash, **cash and cash equivalents at the end of 2025 were \$10.270 billion.**

- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** false positive
- **Diagnostic reviewer root cause:** Validator could not normalize abbreviated billion-dollar units: the model's \$11.488 billion is numerically equivalent to the evidence's \$11,488 million, and its final \$10.270 billion equals the correct answer of \$10,270 million.
- **Verifier verdict:** false positive (confidence: high)
- **Validator flags:** Unsupported number: '\$11.488' does not appear in the evidence

Failure 125 — `SEC\_10K\_multi\_neg\_0581\_03` · — confidence

DETERMINISTIC:ANSWERED UNANSWERABLE

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What is the percentage of total loans that are noninvestment-grade and held for more than 5 years?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** Noninvestment-grade loans originated prior to 2021 totaled:
 - Real estate: \$10,978 million
 - Commercial and industrial: \$825 million
 - Other: \$2,563 million

Total: **\$14,366 million**

$$\frac{14,366}{792,367} \times 100 \approx 1.81\%$$

Therefore, **approximately 1.8% of total retained loans** were noninvestment-grade and held for more than five years.

- **Diagnostic reviewer diagnosis:** **Negative-answer specialist** model error
- **Diagnostic reviewer root cause:** [code:answered_unanswerable] GT is UNANSWERABLE but the response asserts a concrete answer
- **Verifier verdict:** — (confidence: —)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer; Unsupported number: '1.8%' does not appear in the evidence

Failure 126 — `SEC\_10K\_multi\_1403\_00` · HIGH confidence

UNSUPPORTED NUMBER

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Cross-Reference
- **Question:** What is the total number of derivatives qualifying as guarantees, and how does it compare to the unsettled resale and securities borrowed agreements?
- ✓ **Ground truth:** 1,168, which is less than 136,841.
- ✗ **Model answered:** As of December 31, 2025:
 - **Derivatives qualifying as guarantees:** **\$49.031 billion**
 - **Unsettled resale and securities borrowed agreements:** **\$137.072 billion**

The unsettled agreements were **\$88.041 billion higher**, or approximately **2.8 times** the derivatives qualifying as guarantees.

- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** model error
- **Diagnostic reviewer root cause:** The model read the total dollar-amount column (49,031 and 137,072, expressed as billions) instead of the count column (1,168 and 136,841), so it compared monetary exposure rather than the requested number of agreements.
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Unsupported number: '\$49.031' does not appear in the evidence

Failure 127 — `SEC\_10K\_multi\_neg\_0767\_03` · — confidence

DETERMINISTIC:ANSWERED UNANSWERABLE

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What is the volatility of natural gas commodity forward?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** The commodity volatility ranged from **2% to 36%**, with a weighted average of **6%**.
- **Diagnostic reviewer diagnosis:** **Negative-answer specialist** model error
- **Diagnostic reviewer root cause:** [code:answered_unanswerable] GT is UNANSWERABLE but the response asserts a concrete answer
- **Verifier verdict:** — (confidence: —)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 128 — `SEC\_10K\_multi\_0404\_00` · HIGH confidence

UNIT/FORMAT MISMATCH

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Cross-Reference
- **Question:** What is the total assets under management in North America compared to the total international assets under management?
- ✓ **Ground truth:** 3,582 vs 1,209
- ✗ **Model answered:** As of December 31, 2025, assets under management were **\$3.582 trillion in North America**, compared with **\$1.209 trillion internationally**. North America had **\$2.373 trillion more** in assets under management.
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** model error
- **Diagnostic reviewer root cause:** [verifier] The model fabricated information not present in the evidence, specifically the unit 'trillion' and the date 'December 31, 2025'.
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Unsupported number: '\$3.582' does not appear in the evidence

Failure 129 — `SEC\_10K\_multi\_neg\_1265\_01` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** How many business combinations occurred in 2025 that resulted in goodwill?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** Zero business combinations occurred in 2025 that resulted in goodwill.
- **Diagnostic reviewer diagnosis:** **Negative-answer specialist** model error

- **Diagnostic reviewer root cause:** The model inferred that no business combinations resulted in goodwill from the absence of a visible 2025 value, but the evidence is truncated immediately after the 'Business combinations' row label and does not provide the relevant figure or count.
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 130 — `SEC\_10K\_multi\_neg\_0003\_06` · HIGH confidence

VALIDATOR OVER-FIRED

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What is the total cost of implementing new tariffs on imports to the U.S.?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** The total cost is not quantified. Apple states that the impact of the new U.S. tariffs remains uncertain and depends on their scope, duration, future changes, and retaliatory measures.
- **Diagnostic reviewer diagnosis:** **Negative-answer specialist** false positive
- **Diagnostic reviewer root cause:** The model correctly recognized that the evidence provides no quantified total cost and explicitly stated that the impact remains uncertain; the validator misclassified this explanatory refusal as a hallucinated answer rather than treating it as equivalent to UNANSWERABLE.
- **Verifier verdict:** false positive (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 131 — `SEC\_10K\_multi\_neg\_0607\_03` · — confidence

DETERMINISTIC:ANSWERED UNANSWERABLE

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What is the total value of the All other category as of December 31, 2025?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** The "All other commissions and fees" category totaled **\$2.049 billion** for 2025.
- **Diagnostic reviewer diagnosis:** **Negative-answer specialist** model error
- **Diagnostic reviewer root cause:** [code:answered_unanswerable] GT is UNANSWERABLE but the response asserts a concrete answer
- **Verifier verdict:** — (confidence: —)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer; Unsupported number: '\$2.049' does not appear in the evidence

Failure 132 — `SEC\_10K\_multi\_neg\_1614\_02` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What is the name of the attorney who certified the report?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** The NVIDIA report appoints **Jen-Hsun Huang and Colette M. Kress** as attorneys-in-fact.
- **Diagnostic reviewer diagnosis:** **Negative-answer specialist** model error

- **Diagnostic reviewer root cause:** The model used an unrelated NVIDIA filing or signature block instead of the Coca-Cola evidence, hallucinating NVIDIA officers as attorneys-in-fact even though the passage does not identify an attorney who certified the report.
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 133 — `SEC\_10K\_multi\_0861\_01` · HIGH confidence

UNSUPPORTED NUMBER

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Precision
- **Question:** What is the value of Short-term borrowings?
- ✓ **Ground truth:** 32.3
- ✗ **Model answered:** Short-term borrowings were **\$0 million** as of December 31, 2025 and 2024.
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** model error
- **Diagnostic reviewer root cause:** Model read the dash in an adjacent classification column as the total value instead of using the Short-term borrowings row's reported value of \$32.3 million.
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Unsupported number: '\$0' does not appear in the evidence

Failure 134 — `SEC\_10K\_multi\_neg\_0539\_04` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What is the credit rating of J.P. Morgan Securities plc according to Standard & Poor's?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** J.P. Morgan Securities plc is rated **AA-** for long-term issuer credit and **A-1+** for short-term issuer credit by Standard & Poor's, with a **Stable** outlook.
- **Diagnostic reviewer diagnosis:** **Negative-answer specialist** model error
- **Diagnostic reviewer root cause:** The model hallucinated specific Standard & Poor's ratings, likely relying on outside knowledge or an assumed table continuation, instead of recognizing that the provided evidence contains no S&P rating values for J.P. Morgan Securities plc.
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 135 — `SEC\_10K\_multi\_1113\_00` · HIGH confidence

UNSUPPORTED NUMBER

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Precision
- **Question:** What is the total loan balance at December 31, 2024?
- ✓ **Ground truth:** 1,347,988
- ✗ **Model answered:** The total retained loan balance at December 31, 2024 was **\$1.300 trillion (\$1,299.590 billion)**.
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** model error
- **Diagnostic reviewer root cause:** Model read the 'Retained' row (\$1,299,590 million) instead of the overall 'Total' row (\$1,347,988 million) for December 31, 2024.

- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Unsupported number: '\$1.300' does not appear in the evidence

Failure 136 — `SEC\_10K\_multi\_1525\_01` · HIGH confidence

UNSUPPORTED NUMBER

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Cross-Reference
- **Question:** How do the assets under management (AUM) and the assets under custody (AUC) relate to the client assets?
- ✓ **Ground truth:** Both AUM and AUC represent assets held on behalf of clients, but AUM includes assets managed by AWM, while AUC includes assets held directly or indirectly under safekeeping.
- ✗ **Model answered:** Client assets are the broader measure:
- **AUM:** Assets managed by Asset & Wealth Management on behalf of clients.
- **AUC:** Assets held on clients' behalf under custody, safekeeping, and servicing arrangements.
- **Client assets:** AUM plus assets in custody, brokerage, administration, and deposit accounts.

For AWM at December 31, 2025, **\$7.118 trillion of client assets** consisted of **\$4.791 trillion of AUM** and **\$2.327 trillion in custody, brokerage, administration, and deposits**. Firmwide AUC was **\$41.172 trillion**, so it is not equivalent to, or wholly included in, AWM client assets.

- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** model error
- **Diagnostic reviewer root cause:** [verifier] The model correctly explained the relationship based on the definitions but then included specific numerical data (e.g., '\$7.118 trillion of client assets') which is not present in the provided evidence, constituting a hallucination.
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Unsupported number: '\$2.327' does not appear in the evidence

Failure 137 — `SEC\_10K\_multi\_0084\_01` · HIGH confidence

VALIDATOR OVER-FIRED

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Cross-Reference
- **Question:** What is the total revenue for the United States and Taiwan for fiscal year 2026, and how does it compare to the total revenue for the same regions in fiscal year 2025?
- ✓ **Ground truth:** The total revenue for the United States for fiscal year 2026 is \$149,617 million, and for Taiwan is \$42,345 million. For fiscal year 2025, the total revenue for the United States is \$77,482 million, and for Taiwan is \$23,600 million.
- ✗ **Model answered:** - **Fiscal 2026:** United States (\$149,617 million) + Taiwan (\$42,345 million) = **\$191,962 million**
- **Fiscal 2025:** United States (\$77,482 million) + Taiwan (\$23,600 million) = **\$101,082 million**

The combined revenue increased by **\$90,880 million**, or approximately **89.9%**, in fiscal 2026.

- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** false positive
- **Diagnostic reviewer root cause:** Validator failed to recognize correctly derived arithmetic totals because \$191,962 million does not appear verbatim in the evidence; it equals \$149,617 million plus \$42,345 million. The model also correctly calculated the 2025 total and year-over-year increase.
- **Verifier verdict:** false positive (confidence: high)
- **Validator flags:** Unsupported number: '\$191,962' does not appear in the evidence

Failure 138 — `SEC\_10K\_multi\_neg\_1265\_04` · — confidence

DETERMINISTIC:ANSWERED UNANSWERABLE

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What is the total amount of goodwill allocated to the Corporate segment in 2025 that was not allocated in 2024?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** Corporate goodwill increased by **\$30 million** in 2025, from \$692 million in 2024 to \$722 million.
- **Diagnostic reviewer diagnosis:** **Negative-answer specialist** model error
- **Diagnostic reviewer root cause:** [code:answered_unanswerable] GT is UNANSWERABLE but the response asserts a concrete answer
- **Verifier verdict:** — (confidence: —)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 139 — `SEC\_10K\_multi\_1039\_00` · HIGH confidence

VALIDATOR OVER-FIRED

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Precision
- **Question:** What is the operating lease income for the year ended December 31, 2023?
- ✓ **Ground truth:** 2,843
- ✗ **Model answered:** The operating lease income for 2023 was **\$2.843 billion**.
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** false positive
- **Diagnostic reviewer root cause:** Validator failed to recognize unit conversion: the table reports 2,843 in millions, which equals \$2.843 billion.
- **Verifier verdict:** false positive (confidence: high)
- **Validator flags:** Unsupported number: '\$2.843' does not appear in the evidence

Failure 140 — `SEC\_10K\_multi\_neg\_1609\_04` · HIGH confidence

BAD QUESTION

- **Category:** Negative (hallucination traps)
- **Question:** What is the name of the trustee for the notes issued by the Company?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** Deutsche Bank Trust Company Americas.
- **Diagnostic reviewer diagnosis:** **Negative-answer specialist** bad test
- **Diagnostic reviewer root cause:** The model inferred that the note forms listed in the exhibit index were governed by the nearby Company indenture and therefore used that indenture's named trustee, Deutsche Bank Trust Company Americas, rather than recognizing that the evidence does not explicitly tie the unspecified notes to that indenture.
- **Verifier verdict:** bad test (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 141 — `SEC\_10K\_multi\_1158\_00` · HIGH confidence

TABLE MISREAD

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Cross-Reference
- **Question:** What is the total amount of retained loans for the year 2022, and how does it compare to the total retained loans for the year 2023?
- ✓ **Ground truth:** The total retained loans for 2022 is \$19,190 million, and for 2023 is \$24,287 million.
- ✗ **Model answered:** Total retained loans were **\$8.967 billion for 2022**, compared with **\$18.794 billion for 2023**. Thus, 2022 was **\$9.827 billion lower**, a decrease of approximately **52.3%**.
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** model error
- **Diagnostic reviewer root cause:** The model extracted figures from the wrong table, row, or reporting context instead of reading the 'Total retained loans' row under the 2022 and 2023 origination-year columns, which show \$19,190 million and \$24,287 million.
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Unsupported number: '\$18.794' does not appear in the evidence

Failure 142 — `SEC\_10K\_multi\_neg\_0003\_00` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What is the total number of countries where Apple's suppliers have manufacturing facilities?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** Apple identifies supplier manufacturing facilities in **7 countries/regions**: China mainland, India, Japan, South Korea, Taiwan, Vietnam, and the United States.
- **Diagnostic reviewer diagnosis:** **Negative-answer specialist** model error
- **Diagnostic reviewer root cause:** The model treated the six primarily named manufacturing countries/regions plus the U.S. as an exhaustive list of all countries containing supplier manufacturing facilities. The passage only identifies where a significant majority of manufacturing occurs and notes sourcing from U.S. partners; it does not give a complete country count.
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 143 — `SEC\_10K\_multi\_1568\_01` · HIGH confidence

BAD QUESTION

- **Category:** Cross-Reference
- **Question:** What is the estimated potential aggregate remaining incremental tax and interest liability for the tax years 2010 through 2025, and how does it compare to the total amount of unused backup lines of credit?
- ✓ **Ground truth:** \$14 billion
- ✗ **Model answered:** The estimated potential aggregate remaining incremental tax and interest liability is approximately **\$14 billion**. This exceeds the **\$6.2 billion** in unused backup lines of credit by about **\$7.8 billion**, or roughly **2.3 times** the available backup credit.
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** bad test
- **Diagnostic reviewer root cause:** The validator treated the correctly derived 2.3-times comparison as unsupported because that exact number is not stated verbatim in the evidence; it is calculated from \$14 billion divided by \$6.2 billion. The \$7.8 billion difference is likewise valid arithmetic.
- **Verifier verdict:** bad test (confidence: high)

- **Validator flags:** Unsupported number: '2.3' does not appear in the evidence

Failure 144 — `SEC\_10K\_multi\_0803\_00` · HIGH confidence

WRONG PERIOD

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Precision
- **Question:** What is the total number of mortgage-backed securities?
- ✓ **Ground truth:** 775
- ✗ **Model answered:** As of December 31, 2025, total mortgage-backed securities classified as trading assets were **\$162.085 billion**.
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** model error
- **Diagnostic reviewer root cause:** Model hallucinated or retrieved a figure from the wrong period/table instead of reading the 'Total mortgage-backed securities' row, whose total is 775.
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Unsupported number: '\$162.085' does not appear in the evidence

Failure 145 — `SEC\_10K\_multi\_0943\_00` · HIGH confidence

TABLE MISREAD

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Precision
- **Question:** What is the total interest rate contracts?
- ✓ **Ground truth:** 256,851
- ✗ **Model answered:** Total interest rate contracts were **\$29.536 trillion in 2025** and **\$29.660 trillion in 2024** (notional amounts).
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** model error
- **Diagnostic reviewer root cause:** The model retrieved figures from the wrong table or disclosure, giving unrelated 2025 and 2024 notional amounts instead of extracting the evidence row's total interest rate contracts value of 256,851.
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Unsupported number: '\$29.536' does not appear in the evidence

Failure 146 — `SEC\_10K\_multi\_0029\_01` · HIGH confidence

UNIT/FORMAT MISMATCH

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Precision
- **Question:** What is the total net sales for the Americas in 2025?
- ✓ **Ground truth:** 178,353
- ✗ **Model answered:** The Americas' total net sales in 2025 were **\$178.353 billion**.
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** false positive
- **Diagnostic reviewer root cause:** The model correctly converted the table's \$178,353 million into \$178.353 billion, but the validator performed a literal number match and could not recognize the unit conversion.
- **Verifier verdict:** false positive (confidence: high)
- **Validator flags:** Unsupported number: '\$178.353' does not appear in the evidence

Failure 147 — `SEC\_10K\_multi\_neg\_0322\_02` · — confidence

DETERMINISTIC:ANSWERED UNANSWERABLE

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What is the percentage of revenue attributed to derivatives in the Fixed Income Markets?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** The filing does not separately disclose derivatives revenue. If principal transactions revenue is used as a proxy, Fixed Income Markets' 2025 percentage was approximately **54.7%**:

$$\frac{\$12.327 \text{ billion}}{\$22.532 \text{ billion}} \times 100 \approx 54.7\%$$

Principal transactions also include non-derivative financial instruments and commodities.

- **Diagnostic reviewer diagnosis:** **Negative-answer specialist** model error
- **Diagnostic reviewer root cause:** [code:answered_unanswerable] GT is UNANSWERABLE but the response asserts a concrete answer
- **Verifier verdict:** — (confidence: —)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer; Unsupported number: '54.7%' does not appear in the evidence

Failure 148 — `SEC\_10K\_multi\_1210\_00` · HIGH confidence

UNSUPPORTED NUMBER

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Precision
- **Question:** What is the total principal amount outstanding?
- ✓ **Ground truth:** 306,687
- ✗ **Model answered:** The total principal amount outstanding was **\$306.687 billion as of December 31, 2025**, compared with **\$266.202 billion in 2024**.
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** model error
- **Diagnostic reviewer root cause:** [verifier] The model's answer contains fabricated information; it states the date is "December 31, 2025" and includes a comparison figure of "\$266.202 billion in 2024", neither of which is supported by the evidence which is dated "December 31, 2024".
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Unsupported number: '\$306.687' does not appear in the evidence

Failure 149 — `SEC\_10K\_multi\_neg\_0322\_04` · HIGH confidence

CERTAINTY MISMATCH

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What is the name of the specific form number for the hedging activities?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** The specific form is **Form 10-K**.
- **Diagnostic reviewer diagnosis:** **Negative-answer specialist** model error
- **Diagnostic reviewer root cause:** The model mistook the document's filing type shown in the footer (Form 10-K) for a specific form associated with hedging activities, even though the passage names no such form.

- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 150 — `SEC\_10K\_multi\_1286\_01` · HIGH confidence

UNIT/FORMAT MISMATCH

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Cross-Reference
- **Question:** What is the estimated future amortization expense for the year 2030, and how does it compare to the gross carrying values of other intangible assets?
- ✓ **Ground truth:** The estimated future amortization expense for 2030 is \$100 million, which is less than the gross carrying values of \$3.5 billion as of December 31, 2025.
- ✗ **Model answered:** The estimated future amortization expense for 2030 is **\$100 million**. This is substantially below the gross carrying values of other intangible assets: **\$3.5 billion in 2025** and **\$3.8 billion in 2024**-about **2.9%** and **2.6%**, respectively.
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** false positive
- **Diagnostic reviewer root cause:** The model correctly answered \$100 million and added valid comparisons to both disclosed gross carrying values. The validator could not recognize 2.9% and 2.6% as arithmetic derivations of \$100 million divided by \$3.5 billion and \$3.8 billion, respectively, because those percentages were not stated verbatim.
- **Verifier verdict:** false positive (confidence: high)
- **Validator flags:** Unsupported number: '2.9%' does not appear in the evidence

Failure 151 — `SEC\_10K\_multi\_1591\_00` · HIGH confidence

VALIDATOR OVER-FIRED

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Cross-Reference
- **Question:** What is the total notional value of derivatives related to the Company's foreign currency economic hedges as of December 31, 2025, and how does it compare to the value as of December 31, 2024?
- ✓ **Ground truth:** The total notional value of derivatives related to the Company's foreign currency economic hedges was \$9,744 million as of December 31, 2025, and \$8,620 million as of December 31, 2024.
- ✗ **Model answered:** The total notional value was **\$9.744 billion** as of December 31, 2025, compared with **\$8.620 billion** as of December 31, 2024-an increase of **\$1.124 billion**, or approximately **13%**.
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** false positive
- **Diagnostic reviewer root cause:** Validator failed to recognize equivalent unit conversion: \$9,744 million equals \$9.744 billion, and \$8,620 million equals \$8.620 billion. The model also correctly calculated the \$1.124 billion (about 13%) increase.
- **Verifier verdict:** false positive (confidence: high)
- **Validator flags:** Unsupported number: '\$9.744' does not appear in the evidence

Failure 152 — `SEC\_10K\_multi\_0487\_00` · — confidence

DETERMINISTIC:GT NUMBER CONFLICT

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Precision
- **Question:** What is the amount of the allowance for credit losses in 2024?

- ✓ **Ground truth:** 541 million
- ✗ **Model answered:** The allowance for credit losses in 2024 was **\$26.8 billion**.
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** model error
- **Diagnostic reviewer root cause:** [code:gt_number_conflict] Short numeric GT; response asserts a different quantity
- **Verifier verdict:** — (confidence: —)
- **Validator flags:** Unsupported number: '\$26.8' does not appear in the evidence

Failure 153 — `SEC\_10K\_multi\_neg\_0080\_04` · HIGH confidence

VALIDATOR OVER-FIRED

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What is the percentage of NVIDIA's operating results that could be impacted by the outcome of the legal proceedings?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** NVIDIA did not quantify a percentage. It stated that an unfavorable outcome could have a **material adverse impact** on operating results, but any possible loss or range of loss could not be reasonably estimated.
- **Diagnostic reviewer diagnosis:** **Negative-answer specialist** false positive
- **Diagnostic reviewer root cause:** The model correctly recognized that the filing provides no percentage and explicitly explained that the impact or loss cannot be reasonably estimated. The validator failed to treat this explanatory refusal as equivalent to UNANSWERABLE.
- **Verifier verdict:** false positive (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 154 — `SEC\_10K\_multi\_1320\_00` · — confidence

DETERMINISTIC:GT NUMBER CONFLICT

- **Severity:** HIGH — an incorrect quantitative result
- **Category:** Precision
- **Question:** What is the fixed rate for Series DD?
- ✓ **Ground truth:** 169,625
- ✗ **Model answered:** The fixed rate for Series DD is **5.75%**.
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** model error
- **Diagnostic reviewer root cause:** [code:gt_number_conflict] Short numeric GT; response asserts a different quantity
- **Verifier verdict:** — (confidence: —)
- **Validator flags:** (none)

Failure 155 — `SEC\_10K\_multi\_0259\_00` · HIGH confidence

UNSUPPORTED NUMBER

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Precision
- **Question:** What is the total noninterest expense?
- ✓ **Ground truth:** 95,640
- ✗ **Model answered:** Total noninterest expense was **\$95.64 billion in 2025**.

- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** model error
- **Diagnostic reviewer root cause:** [verifier] The model invented information not present in the evidence by stating the year was '2025' and assuming the units were in millions to convert '95,640' to '\$95.64 billion'.
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Unsupported number: '\$95.64' does not appear in the evidence

Failure 156 — `SEC\_10K\_multi\_1419\_01` · HIGH confidence

UNIT/FORMAT MISMATCH

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Cross-Reference
- **Question:** What is the maximum exposure of stable value contracts with contractually limited exposure, and how does it compare to the notional amount?
- ✓ **Ground truth:** 1,312 million, which is a substantially lower percentage of the notional amount
- ✗ **Model answered:** As of December 31, 2025, the maximum exposure was **\$1.312 billion**, compared with a **\$35.462 billion notional amount-about 3.7% of notional**. In 2024, it was **\$1.740 billion** versus **\$32.939 billion**, or about **5.3%**.
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** false positive
- **Diagnostic reviewer root cause:** Validator could not recognize the unit conversion from 1,312 million in the evidence to \$1.312 billion in the response. The response correctly compared it with the \$35.462 billion notional amount and accurately calculated approximately 3.7%.
- **Verifier verdict:** false positive (confidence: high)
- **Validator flags:** Unsupported number: '\$1.312' does not appear in the evidence

Failure 157 — `SEC\_10K\_multi\_neg\_0581\_00` · — confidence

DETERMINISTIC:ANSWERED UNANSWERABLE

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What is the total value of loans held-for-sale?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** As of December 31, 2025, total loans held-for-sale were **\$13.84 billion**.
- **Diagnostic reviewer diagnosis:** **Negative-answer specialist** model error
- **Diagnostic reviewer root cause:** [code:answered_unanswerable] GT is UNANSWERABLE but the response asserts a concrete answer
- **Verifier verdict:** — (confidence: —)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer; Unsupported number: '\$13.84' does not appear in the evidence

Failure 158 — `SEC\_10K\_multi\_0080\_01` · HIGH confidence

UNIT/FORMAT MISMATCH

- **Category:** Cross-Reference
- **Question:** What is the income tax expense for the year ended January 25, 2026, and how does it relate to the U.S. federal statutory tax rate?
- ✓ **Ground truth:** The income tax expense for the year ended January 25, 2026, is \$21,383, which is \$8,257 less than the \$29,704 computed by applying the U.S. federal statutory tax rate of 21.0% to income before income taxes.

- **✗ Model answered:** For the year ended January 25, 2026, income tax expense was **\$21.383 billion**, representing an effective tax rate of **15.1%**. This was **5.9 percentage points below** the U.S. federal statutory tax rate of **21.0%**, primarily due to benefits from foreign-derived deduction eligible income, stock-based compensation, lower-tax foreign earnings, and tax credits.
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** bad test
- **Diagnostic reviewer root cause:** The validator required an explicitly stated number and did not accept the derived 5.9-percentage-point difference between the disclosed 21.0% statutory rate and 15.1% effective rate. The model also correctly converted \$21,383 million to \$21.383 billion.
- **Verifier verdict:** bad test (confidence: high)
- **Validator flags:** Unsupported number: '5.9' does not appear in the evidence

Failure 159 — `SEC\_10K\_multi\_neg\_0781\_01` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** How many U.S. government agency securities are listed?
- **✓ Ground truth:** UNANSWERABLE
- **✗ Model answered:** None. U.S. government agency securities are excluded from the tables.
- **Diagnostic reviewer diagnosis:** **Negative-answer specialist** model error
- **Diagnostic reviewer root cause:** [verifier] The model incorrectly stated that 'U.S. government agency securities are excluded from the tables' when the evidence explicitly lists a row for 'U.S. GSEs and government agencies'.
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 160 — `SEC\_10K\_multi\_1601\_00` · HIGH confidence

UNIT/FORMAT MISMATCH

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Cross-Reference
- **Question:** What is the total amount of assets measured at fair value on a recurring basis at December 31, 2025, according to Note 110?
- **✓ Ground truth:** 4,451
- **✗ Model answered:** The total assets measured at fair value on a recurring basis were **\$4.451 billion** as of December 31, 2025.
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** false positive
- **Diagnostic reviewer root cause:** The model correctly converted the table's \$4,451 million total into \$4.451 billion, but the validator relied on exact numeric-string matching and did not recognize the unit conversion.
- **Verifier verdict:** false positive (confidence: high)
- **Validator flags:** Unsupported number: '\$4.451' does not appear in the evidence

Failure 161 — `SEC\_10K\_multi\_0378\_00` · HIGH confidence

WRONG ENTITY

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Precision
- **Question:** What is the revenue for 2023?

- ✓ **Ground truth:** 11,826
- ✗ **Model answered:** Apple's 2023 total net sales (revenue) were **\$383.285 billion**.
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** model error
- **Diagnostic reviewer root cause:** Model answered from the wrong company and financial statement, recalling Apple's fiscal 2023 total net sales instead of reading JPMorgan Asset & Wealth Management's 2023 asset management fee revenue from the provided table.
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Unsupported number: '\$383.285' does not appear in the evidence

Failure 162 — `SEC\_10K\_multi\_0086\_01` · HIGH confidence

BAD QUESTION

- **Category:** Cause & Effect
- **Question:** What is the result of operating leases for our offices and data centers?
- ✓ **Ground truth:** Our lease obligations primarily consist of operating leases.
- ✗ **Model answered:** As of January 25, 2026, operating leases for offices and data centers resulted in:
 - **\$3.624 billion** of future minimum lease obligations
 - **\$2.944 billion** present value after \$680 million of imputed interest
 - **\$372 million** of short-term liabilities
 - **\$2.572 billion** of long-term operating lease liabilities

The leases had a weighted-average remaining term of **8.8 years** and a weighted-average discount rate of **4.38%**.

- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** bad test
- **Diagnostic reviewer root cause:** The model interpreted the awkward phrase "result of operating leases" as asking for the quantitative lease balances rather than what the lease obligations primarily consist of. It also converted \$3,624 million to the equivalent \$3.624 billion, which the validator failed to recognize.
- **Verifier verdict:** bad test (confidence: high)
- **Validator flags:** Unsupported number: '\$3.624' does not appear in the evidence

Failure 163 — `SEC\_10K\_multi\_neg\_0744\_01` · — confidence

DETERMINISTIC:ANSWERED UNANSWERABLE

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What is the total amount of derivative netting adjustments?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** The total derivative netting adjustments were **\$1.357 billion** in 2025, comprising **\$403 million for assets** and **\$954 million for liabilities**.
- **Diagnostic reviewer diagnosis:** **Negative-answer specialist** model error
- **Diagnostic reviewer root cause:** [code:answered_unanswerable] GT is UNANSWERABLE but the response asserts a concrete answer
- **Verifier verdict:** — (confidence: —)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer; Unsupported number: '\$1.357' does not appear in the evidence

Failure 164 — `SEC\_10K\_multi\_neg\_0744\_04` · — confidence

DETERMINISTIC:ANSWERED UNANSWERABLE

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What is the total amount of beneficial interests issued by consolidated VIEs at Level 1 of the fair value hierarchy?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** The Level 1 amount for beneficial interests issued by consolidated VIEs was \$0.
- **Diagnostic reviewer diagnosis:** **Negative-answer specialist** model error
- **Diagnostic reviewer root cause:** [code:answered_unanswerable] GT is UNANSWERABLE but the response asserts a concrete answer
- **Verifier verdict:** — (confidence: —)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer; Unsupported number: '\$0' does not appear in the evidence

Failure 165 — `SEC\_10K\_multi\_0023\_01` · HIGH confidence

UNSUPPORTED NUMBER

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Precision
- **Question:** What is the value of the total?
- ✓ **Ground truth:** 132,420
- ✗ **Model answered:** The total net sales for 2025 were **\$416.161 billion** (\$416,161 million).
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** model error
- **Diagnostic reviewer root cause:** Model answered from the wrong financial table or metric, giving 2025 net sales instead of reading the evidence table's Total fair value of 132,420.
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Unsupported number: '\$416.161' does not appear in the evidence

Failure 166 — `SEC\_10K\_multi\_neg\_1274\_03` · — confidence

DETERMINISTIC:ANSWERED UNANSWERABLE

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What is the percentage change in valuation due to inputs and assumptions?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** From 2024 to 2025, changes in valuation due to inputs and assumptions fell from **\$762 million to \$48 million**:

$$\left[\frac{48-762}{762} \right] \times 100 \approx -93.7\%$$

Thus, the valuation change **decreased by approximately 93.7%**.

- **Diagnostic reviewer diagnosis:** **Negative-answer specialist** model error
- **Diagnostic reviewer root cause:** [code:answered_unanswerable] GT is UNANSWERABLE but the response asserts a concrete answer
- **Verifier verdict:** — (confidence: —)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 167 — `SEC\_10K\_multi\_1062\_00` · HIGH confidence

VALIDATOR OVER-FIRED

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Cross-Reference
- **Question:** What is the expected long-term rate of return on plan assets for the year ended December 31, 2025, and how does it compare to the rate for 2024?
- ✓ **Ground truth:** 5.44% for 2025 and 6.15% for 2024.
- ✗ **Model answered:** The expected long-term rate of return on plan assets was **5.44% in 2025**, down from **6.15% in 2024**, a decrease of **0.71 percentage points**.
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** false positive
- **Diagnostic reviewer root cause:** The model correctly quoted both rates and additionally computed their difference ($6.15\% - 5.44\% = 0.71$ percentage points). The validator only accepts numbers explicitly present in the evidence and could not validate the derived comparison.
- **Verifier verdict:** false positive (confidence: high)
- **Validator flags:** Unsupported number: '0.71' does not appear in the evidence

Failure 168 — `SEC\_10K\_multi\_neg\_0744\_02` · HIGH confidence

BAD QUESTION

- **Category:** Negative (hallucination traps)
- **Question:** What is the total amount of long-term debt measured at fair value on a recurring basis?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** The total long-term debt measured at fair value on a recurring basis was **\$134.559 billion** as of December 31, 2025.
- **Diagnostic reviewer diagnosis:** **Negative-answer specialist** bad test
- **Diagnostic reviewer root cause:** The model correctly read the long-term debt total of \$134,559 million from the table and converted it to \$134.559 billion, but the benchmark incorrectly labels the question unanswerable and the validator failed to recognize the unit conversion.
- **Verifier verdict:** bad test (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer; Unsupported number: '\$134.559' does not appear in the evidence

Failure 169 — `SEC\_10K\_multi\_neg\_0034\_01` · — confidence

DETERMINISTIC:ANSWERED UNANSWERABLE

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What is the average interest rate of the notes issued by Apple Inc. in 2025?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** Apple's 2025 notes had fixed rates ranging from **4.00% to 4.75%**, giving a simple average of **4.375%**.
- **Diagnostic reviewer diagnosis:** **Negative-answer specialist** model error
- **Diagnostic reviewer root cause:** [code:answered_unanswerable] GT is UNANSWERABLE but the response asserts a concrete answer
- **Verifier verdict:** — (confidence: —)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer; Unsupported number: '4.375%' does not appear in the evidence

Failure 170 — `SEC\_10K\_multi\_neg\_1609\_03` · — confidence

DETERMINISTIC:ANSWERED UNANSWERABLE

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What is the total value of the notes issued by the Company with a maturity date of 2030?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** The notes maturing in 2030 total **\$1.5 billion**.
- **Diagnostic reviewer diagnosis:** **Negative-answer specialist** model error
- **Diagnostic reviewer root cause:** [code:answered_unanswerable] GT is UNANSWERABLE but the response asserts a concrete answer
- **Verifier verdict:** — (confidence: —)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer; Unsupported number: '\$1.5' does not appear in the evidence

Failure 171 — `SEC\_10K\_multi\_1471\_00` · HIGH confidence

VALIDATOR OVER-FIRED

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Cross-Reference
- **Question:** What is the total value of lease right-of-use assets and the related tax-oriented investment liabilities?
- ✓ **Ground truth:** \$683 million and \$669 million
- ✗ **Model answered:** The total is **\$1.352 billion**: \$683 million in lease right-of-use assets plus \$669 million in tax-oriented investment liabilities.
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** false positive
- **Diagnostic reviewer root cause:** The model correctly quoted both figures and summed them to \$1.352 billion; the validator only checked whether the aggregate appeared verbatim in the evidence and did not recognize the valid arithmetic derivation.
- **Verifier verdict:** false positive (confidence: high)
- **Validator flags:** Unsupported number: '\$1.352' does not appear in the evidence

Failure 172 — `SEC\_10K\_multi\_0822\_00` · HIGH confidence

UNIT/FORMAT MISMATCH

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Cross-Reference
- **Question:** What is the total realized/unrealized gains/(losses) for the year ended December 31, 2023, and what is the fair value at January 1, 2023?
- ✓ **Ground truth:** 738, 27,726
- ✗ **Model answered:** For **deposits**, total realized/unrealized gains/(losses) were **\$95 million**, and fair value at January 1, 2023 was **\$2.162 billion**.
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** model error
- **Diagnostic reviewer root cause:** Model read the values from the deposits row or another nearby table entry instead of the long-term debt row for 2023; the relevant figures are \$738 million in gains and \$27.726 billion fair value at January 1, 2023.
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Unsupported number: '\$2.162' does not appear in the evidence

Failure 173 — `SEC\_10K\_multi\_0646\_00` · HIGH confidence

UNSUPPORTED NUMBER

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Precision
- **Question:** What is the average retained loans?
- ✓ **Ground truth:** 371,238
- ✗ **Model answered:** The average retained residential real estate FDM loans for 2023-2025 were approximately **\$455.3 million**:

\[
(\$1{,}024M+\$206M+\$136M)\div3=\$455.3M
\]

- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** model error
- **Diagnostic reviewer root cause:** Model used unrelated retained residential real estate FDM loan figures and computed a three-year average instead of reading the stated 'Retained loans, average' value of 371,238 from the table.
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Unsupported number: '\$455.3' does not appear in the evidence

Failure 174 — `SEC\_10K\_multi\_neg\_0039\_00` · — confidence

DETERMINISTIC:ANSWERED UNANSWERABLE

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What is the total number of employees at NVIDIA Corporation?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** As of December 31, 2025, NVIDIA Corporation had approximately **65,900 employees**.
- **Diagnostic reviewer diagnosis:** **Negative-answer specialist** model error
- **Diagnostic reviewer root cause:** [code:answered_unanswerable] GT is UNANSWERABLE but the response asserts a concrete answer
- **Verifier verdict:** — (confidence: —)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 175 — `SEC\_10K\_multi\_1290\_01` · — confidence

DETERMINISTIC:DIRECTION CONFLICT

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Conditional
- **Question:** If the lease term is 7 years and the estimated useful life of leasehold improvements is 10 years, does the Firm depreciate leasehold improvements over the lease term?
- ✓ **Ground truth:** No
- ✗ **Model answered:** Yes. The Firm depreciates leasehold improvements over the lesser of the remaining lease term or estimated useful life. Since 7 years is shorter than 10 years, it would depreciate them over the 7-year lease term.
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** model error
- **Diagnostic reviewer root cause:** [code:direction_conflict] GT is 'no' but response leads 'yes'
- **Verifier verdict:** — (confidence: —)
- **Validator flags:** Direction mismatch: GT negative, response affirmative

Failure 176 — `SEC\_10K\_multi\_neg\_0607\_01` · — confidence

DETERMINISTIC:ANSWERED UNANSWERABLE

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What percentage of the total credit exposure is attributed to loans and lending-related commitments?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** Loans and lending-related commitments represent approximately **98.3%** of total Consumer & Retail credit exposure:

\[
\\\$131.710\\text{B} \\div \\\$133.945\\text{B} \\times 100 \\approx 98.3\\%
\\]

- **Diagnostic reviewer diagnosis:** **Negative-answer specialist** model error
- **Diagnostic reviewer root cause:** [code:answered_unanswerable] GT is UNANSWERABLE but the response asserts a concrete answer
- **Verifier verdict:** — (confidence: —)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer; Unsupported number: '98.3%' does not appear in the evidence

Failure 177 — `SEC\_10K\_multi\_neg\_1595\_04` · — confidence

DETERMINISTIC:ANSWERED UNANSWERABLE

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What is the average annual return on the Company's other postretirement benefit plans' assets for the year 2025?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** The 2025 annual return was approximately **13.0%**, calculated as:

\[
\\\$9\\text{ million} \\div \\\$69\\text{ million} \\times 100 \\approx 13.0\\%
\\]

- **Diagnostic reviewer diagnosis:** **Negative-answer specialist** model error
- **Diagnostic reviewer root cause:** [code:answered_unanswerable] GT is UNANSWERABLE but the response asserts a concrete answer
- **Verifier verdict:** — (confidence: —)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 178 — `SEC\_10K\_multi\_1587\_00` · HIGH confidence

UNIT/FORMAT MISMATCH

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Cross-Reference
- **Question:** What is the total amount of proceeds from disposals of businesses, equity method investments, and nonmarketable securities in 2025, and how does it compare to the total in 2024?
- ✓ **Ground truth:** The total amount of proceeds in 2025 is \$3,567 million, which is \$82 million more than the total in 2024, which was \$3,485 million.

- **Model answered:** Proceeds totaled **\$3.567 billion in 2025**, compared with **\$3.485 billion in 2024**-an increase of **\$82 million**, or approximately **2.4%**.
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** false positive
- **Diagnostic reviewer root cause:** Validator could not normalize the unit conversion and formatting: \$3.567 billion is numerically equivalent to the evidence's \$3,567 million. The model's 2024 total, \$82 million increase, and approximately 2.4% increase are also correct.
- **Verifier verdict:** false positive (confidence: high)
- **Validator flags:** Unsupported number: '\$3.567' does not appear in the evidence

Failure 179 — `SEC\_10K\_multi\_neg\_1585\_02` · — confidence

DETERMINISTIC:ANSWERED UNANSWERABLE

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What is the total value of tax credits received from limited partnerships in 2025?
- **Ground truth:** UNANSWERABLE
- **Model answered:** The total tax credits and other income tax benefits received from limited partnerships in 2025 were **\$241 million**.
- **Diagnostic reviewer diagnosis:** **Negative-answer specialist** model error
- **Diagnostic reviewer root cause:** [code:answered_unanswerable] GT is UNANSWERABLE but the response asserts a concrete answer
- **Verifier verdict:** — (confidence: —)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 180 — `SEC\_10K\_multi\_1453\_01` · HIGH confidence

UNSUPPORTED NUMBER

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Precision
- **Question:** What is the amount of cash and due from banks for 2024?
- **Ground truth:** 38
- **Model answered:** Cash and due from banks and deposits with banks at December 31, 2024 was **\$469.317 billion**.
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** model error
- **Diagnostic reviewer root cause:** Model retrieved an unrelated balance-sheet figure and combined cash and due from banks with deposits at banks, instead of reading the 2024 value of \$38 million from the correct row and column.
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Unsupported number: '\$469.317' does not appear in the evidence

Failure 181 — `SEC\_10K\_multi\_0301\_01` · — confidence

DETERMINISTIC:GT NUMBER CONFLICT

- **Severity:** HIGH — an incorrect quantitative result
- **Category:** Precision
- **Question:** What is the net charge-off/(recovery) rate for Auto?
- **Ground truth:** 0.49
- **Model answered:** The Auto net charge-off/(recovery) rate was **0.47% in 2025**.

- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** model error
- **Diagnostic reviewer root cause:** [code:gt_number_conflict] Short numeric GT; response asserts a different quantity
- **Verifier verdict:** — (confidence: —)
- **Validator flags:** (none)

Failure 182 — `SEC\_10K\_multi\_0075\_01` · HIGH confidence

UNIT/FORMAT MISMATCH

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Precision
- **Question:** What was the total of other long-term liabilities as of January 26, 2025?
- ✓ **Ground truth:** 4,245
- ✗ **Model answered:** The total other long-term liabilities as of January 26, 2025 were **\$4.245 billion**.
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** false positive
- **Diagnostic reviewer root cause:** Validator could not normalize equivalent units and formatting: \$4.245 billion equals 4,245 million, the figure shown in the evidence.
- **Verifier verdict:** false positive (confidence: high)
- **Validator flags:** Unsupported number: '\$4.245' does not appear in the evidence

Failure 183 — `SEC\_10K\_multi\_0019\_01` · HIGH confidence

VALIDATOR OVER-FIRED

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Cross-Reference
- **Question:** What is the total net sales for the years ended September 27, 2025, September 28, 2024, and September 30, 2023, and how do they compare to each other?
- ✓ **Ground truth:** \$416,161, \$391,035, and \$383,285
- ✗ **Model answered:** Total net sales were:
 - **2025:** \$416.161 billion
 - **2024:** \$391.035 billion
 - **2023:** \$383.285 billion

Net sales increased by **\$25.126 billion (6.4%)** from 2024 to 2025 and by **\$7.750 billion (2.0%)** from 2023 to 2024. Overall, sales increased **\$32.876 billion (8.6%)** from 2023 to 2025.

- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** false positive
- **Diagnostic reviewer root cause:** Validator failed to recognize a unit conversion: the table reports figures in millions, so \$416,161 million equals \$416.161 billion, with the other values and comparisons also correct.
- **Verifier verdict:** false positive (confidence: high)
- **Validator flags:** Unsupported number: '\$416.161' does not appear in the evidence

Failure 184 — `SEC\_10K\_multi\_0024\_00` · HIGH confidence

UNSUPPORTED NUMBER

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Cross-Reference
- **Question:** What is the total value of U.S. Treasury securities and U.S. agency securities combined?
- ✓ **Ground truth:** $16,150 + 5,431 = 21,581$

- **✗ Model answered:** As of September 27, 2025, their combined **fair value was \$20.968 billion** (\$15.848 billion + \$5.120 billion).
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** model error
- **Diagnostic reviewer root cause:** The model pulled figures from a different reporting date or table and summed \$15.848 billion and \$5.120 billion instead of the evidence's U.S. Treasury and agency values of \$16.150 billion and \$5.431 billion, which total \$21.581 billion.
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Unsupported number: '\$20.968' does not appear in the evidence

Failure 185 — `SEC\_10K\_multi\_1155\_00` · HIGH confidence

UNIT/FORMAT MISMATCH

- **Category:** Cross-Reference
- **Question:** What is the total retained loans for 2025 and 2024, and what percentage of investment-grade loans is that for each year?
- **✓ Ground truth:** The total retained loans for 2025 is \$165,545 and for 2024 is \$162,432. The percentage of investment-grade loans for 2025 is 71.81% and for 2024 is 70.36%.
- **✗ Model answered:** - **2025:** Total retained loans were **\$792.367 billion**, of which **68.32%** were investment-grade.
- **2024:** Total retained loans were **\$690.396 billion**, of which **68.32%** were investment-grade.
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** bad test
- **Diagnostic reviewer root cause:** The model selected the fourth 2025/2024 column pair (\$792,367 and \$690,396, with 68.32%) instead of the first pair used by the ground truth (\$165,545 and \$162,432, with 71.81% and 70.36%). It also inferred billions from unstated table units.
- **Verifier verdict:** bad test (confidence: high)
- **Validator flags:** Unsupported number: '\$792.367' does not appear in the evidence

Failure 186 — `SEC\_10K\_multi\_1591\_01` · HIGH confidence

UNIT/FORMAT MISMATCH

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Cross-Reference
- **Question:** What is the difference between the calculated fair value and the carrying value of the Company's investment in Monster Beverage Corporation as of December 31, 2025?
- **✓ Ground truth:** The difference between the calculated fair value and the carrying value of the Company's investment in Monster Beverage Corporation was \$10,066 million as of December 31, 2025.
- **✗ Model answered:** The difference was **\$10,066 million** (approximately **\$10.1 billion**).
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** false positive
- **Diagnostic reviewer root cause:** The model correctly quoted \$10,066 million and additionally converted it to approximately \$10.1 billion. The validator relied on exact textual number matching and could not recognize the valid rounded unit conversion.
- **Verifier verdict:** false positive (confidence: high)
- **Validator flags:** Unsupported number: '\$10.1' does not appear in the evidence

Failure 187 — `SEC\_10K\_multi\_1160\_01` · HIGH confidence

TABLE MISREAD

- **Category:** Cross-Reference
- **Question:** What is the total retained loans amount in the U.S. and the total loan delinquency amount in the 30-89 days past due category, respectively?

- ✓ **Ground truth:** \$165,545 and 918
- ✗ **Model answered:** As of December 31, 2025:
- **Total retained U.S. loans:** \$626.060 billion
- **Loans 30-89 days past due:** \$2.913 billion
- **Diagnostic reviewer diagnosis:** General diagnostic specialist bad test
- **Diagnostic reviewer root cause:** [verifier] The ground truth is incorrect because the question asks for the 'total retained loans amount in the U.S.', but the ground truth value of '\$165,545' is from the 'Total retained loans' row, which is the sum of U.S. and non-U.S. loans.
- **Verifier verdict:** bad test (confidence: high)
- **Validator flags:** Unsupported number: '\$626.060' does not appear in the evidence

Failure 188 — `SEC\_10K\_multi\_neg\_0071\_04` · — confidence

DETERMINISTIC:ANSWERED UNANSWERABLE

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What is the estimated fair value of non-marketable equity securities as of January 26, 2025?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** The reported carrying value of non-marketable equity securities as of January 26, 2025 was \$3.387 billion.
- **Diagnostic reviewer diagnosis:** Negative-answer specialist model error
- **Diagnostic reviewer root cause:** [code:answered_unanswerable] GT is UNANSWERABLE but the response asserts a concrete answer
- **Verifier verdict:** — (confidence: —)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer; Unsupported number: '\$3.387' does not appear in the evidence

Failure 189 — `SEC\_10K\_multi\_0285\_00` · HIGH confidence

UNIT/FORMAT MISMATCH

- **Category:** Cross-Reference
- **Question:** What is the total revenue for the year ended December 31, 2024, as reported in the selected income statement data, and how does it compare to the year ended December 31, 2025?
- ✓ **Ground truth:** The total revenue for the year ended December 31, 2024, is \$3,387 million, and for the year ended December 31, 2025, it is \$3,669 million.
- ✗ **Model answered:** Total net revenue for 2024 was \$177.556 billion, compared with \$182.447 billion in 2025—an increase of \$4.891 billion, or 3%, in 2025.
- **Diagnostic reviewer diagnosis:** General diagnostic specialist bad test
- **Diagnostic reviewer root cause:** The model pulled JPMorgan's firmwide total net revenue figures from a different financial table instead of the intended Consumer & Community Banking selected-data row.
- **Verifier verdict:** bad test (confidence: high)
- **Validator flags:** Unsupported number: '\$177.556' does not appear in the evidence

Failure 190 — `SEC\_10K\_multi\_0084\_00` · HIGH confidence

VALIDATOR OVER-FIRED

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Cross-Reference

- **Question:** What is the total depreciation and amortization expense for the Compute & Networking segment for fiscal years 2026 and 2025, and how does it compare to the Graphics segment for the same years?
- **Ground truth:** The total depreciation and amortization expense for the Compute & Networking segment for fiscal years 2026 and 2025 is \$1.6 billion and \$732 million, respectively. For the Graphics segment, the total depreciation and amortization expense for fiscal years 2026 and 2025 is \$590 million and \$372 million, respectively.
- **Model answered:** - **Compute & Networking:** \$1.6 billion (2026) + \$732 million (2025) = **\$2.332 billion**
- **Graphics:** \$590 million (2026) + \$372 million (2025) = **\$962 million**

Compute & Networking's total was **\$1.370 billion higher**, about **2.4 times** the Graphics segment's total.

- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** false positive
- **Diagnostic reviewer root cause:** The model quoted the correct annual figures, then correctly computed two-year aggregates and the difference; the validator rejected the derived \$2.332 billion because it does not appear verbatim in the evidence.
- **Verifier verdict:** false positive (confidence: high)
- **Validator flags:** Unsupported number: '\$2.332' does not appear in the evidence

Failure 191 — `SEC\_10K\_multi\_1567\_01` · HIGH confidence

VALIDATOR OVER-FIRED

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Cause & Effect
- **Question:** What is the effect of lower other operating charges on the Company's Corporate segment?
- **Ground truth:** decrease in operating loss
- **Model answered:** Lower other operating charges substantially reduced the Corporate segment's operating loss—from \$5.263 billion in 2024 to \$1.816 billion in 2025—primarily because the prior year included a large fairlife contingent-consideration remeasurement charge.
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** false positive
- **Diagnostic reviewer root cause:** Validator failed to recognize a valid unit conversion: the model converted \$5,263 million and \$1,816 million from the evidence into \$5.263 billion and \$1.816 billion. Its substantive conclusion that lower charges decreased the Corporate segment's operating loss matches the correct answer.
- **Verifier verdict:** false positive (confidence: high)
- **Validator flags:** Unsupported number: '\$5.263' does not appear in the evidence

Failure 192 — `SEC\_10K\_multi\_0932\_01` · HIGH confidence

BAD QUESTION

- **Category:** Precision
- **Question:** What is the total derivative receivables?
- **Ground truth:** 2,609
- **Model answered:** As of December 31, 2025, total derivative receivables were **\$606.382 billion** gross, or **\$57.777 billion** after netting adjustments.
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** bad test
- **Diagnostic reviewer root cause:** The model correctly read the table's total gross derivative receivables of \$606,382 million and converted it to \$606.382 billion; the validator failed to recognize the unit conversion and the reference answer appears to select the adjacent \$2,609 million designated-as-hedges payables value.
- **Verifier verdict:** bad test (confidence: high)
- **Validator flags:** Unsupported number: '\$606.382' does not appear in the evidence

Failure 193 — `SEC\_10K\_multi\_0629\_01` · — confidence

DETERMINISTIC:GT NUMBER CONFLICT

- **Severity:** HIGH — an incorrect quantitative result
- **Category:** Precision
- **Question:** What is the net charge-off/(recovery) rate for the year ended December 31, 2025?
- ✓ **Ground truth:** 0.22
- ✗ **Model answered:** The net charge-off/(recovery) rate for 2025 was **0.29%**.
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** model error
- **Diagnostic reviewer root cause:** [code:gt_number_conflict] Short numeric GT; response asserts a different quantity
- **Verifier verdict:** — (confidence: —)
- **Validator flags:** (none)

Failure 194 — `SEC\_10K\_multi\_0537\_00` · HIGH confidence

UNSUPPORTED NUMBER

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Precision
- **Question:** What is the amount of FHLB advances in 2024?
- ✓ **Ground truth:** 6,000
- ✗ **Model answered:** FHLB advances were **\$29.257 billion as of December 31, 2024**. The 2024 average balance was **\$35.040 billion**.
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** model error
- **Diagnostic reviewer root cause:** [verifier] The model provided figures (\$29.257 billion and \$35.040 billion) that are not present in the evidence, while the ground truth (\$6,000) is explicitly stated in the evidence table as the 'Issuance' for 'FHLB advances' in 2024.
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Unsupported number: '\$29.257' does not appear in the evidence

Failure 195 — `SEC\_10K\_multi\_neg\_0618\_01` · — confidence

DETERMINISTIC:ANSWERED UNANSWERABLE

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What is the dollar amount of investment-grade loans as of December 31, 2025?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** Investment-grade retained loans totaled **\$541.364 billion** (\$541,364 million) as of December 31, 2025.
- **Diagnostic reviewer diagnosis:** **Negative-answer specialist** model error
- **Diagnostic reviewer root cause:** [code:answered_unanswerable] GT is UNANSWERABLE but the response asserts a concrete answer
- **Verifier verdict:** — (confidence: —)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer; Unsupported number: '\$541.364' does not appear in the evidence

Failure 196 — `SEC\_10K\_multi\_neg\_1279\_02` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What is the maximum amount the Firm can advance to investors?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** The Firm can advance only up to the amount covered by the collateral; it may stop payments to investors if the collateral is insufficient.
- **Diagnostic reviewer diagnosis:** Negative-answer specialist model error
- **Diagnostic reviewer root cause:** [verifier] The evidence does not state a maximum amount the firm can advance, so the correct answer is UNANSWERABLE; the model incorrectly inferred a limit from the firm's right to stop payments, which is a fabrication and thus a model error, not a fault in the test.
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 197 — `SEC\_10K\_multi\_1345\_01` · HIGH confidence

UNSUPPORTED NUMBER

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Precision
- **Question:** What is the total other comprehensive income/(loss)?
- ✓ **Ground truth:** 8,997
- ✗ **Model answered:** Apple's total other comprehensive income was **\$1.601 billion in 2025**.
- **Diagnostic reviewer diagnosis:** General diagnostic specialist model error
- **Diagnostic reviewer root cause:** [verifier] The evidence contains a clear row labeled 'Total other comprehensive income/(loss)' with the value '\$ 8,997', which matches the ground truth; the model hallucinated an answer about 'Apple' and '2025', which are not mentioned in the evidence, making this a model error, not a bad test.
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Unsupported number: '\$1.601' does not appear in the evidence

Failure 198 — `SEC\_10K\_multi\_1274\_01` · HIGH confidence

UNSUPPORTED NUMBER

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Cross-Reference
- **Question:** What is the fair value at December 31, and how does it relate to the change in unrealized gains/(losses) included in income?
- ✓ **Ground truth:** \$9,167, which is greater than \$48
- ✗ **Model answered:** The question does not specify the financial instrument. For the Level 3 liabilities at December 31, 2023:
 - Deposits: fair value **\$1,833 million**; unrealized gain/loss included in income for instruments still held: **\$73 million**
 - Short-term borrowings: **\$1,758 million**; **\$14 million**
 - Trading liabilities: **\$37 million**; **\$0**
 - Accounts payable and other liabilities: **\$52 million**; **\$(4) million**

- Long-term debt: **\$27,726 million; \$2,870 million**

The unrealized amount is the portion of the period's fair-value change attributable to instruments still held at year-end; it is not the total ending fair value.

- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** model error
- **Diagnostic reviewer root cause:** The model selected an unrelated Level 3 liabilities table and reported values for several instruments instead of using the MSR row in the provided evidence, where the December 31 fair value is \$9,167 million and the unrealized gain/loss is \$48 million.
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Unsupported number: '\$0' does not appear in the evidence

Failure 199 — `SEC\_10K\_multi\_0019\_00` · HIGH confidence

VALIDATOR OVER-FIRED

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Cross-Reference
- **Question:** What is the maximum one-day loss in fair value of the Company's foreign currency derivative positions due to adverse movements in exchange rates, as of September 27, 2025, and how does it compare to the loss in 2024?
- ✓ **Ground truth:** \$590 million and \$538 million
- ✗ **Model answered:** As of September 27, 2025, the maximum one-day loss was estimated at **\$590 million** with 95% confidence, compared with **\$538 million** as of September 28, 2024—an increase of **\$52 million**, or approximately 9.7%.
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** false positive
- **Diagnostic reviewer root cause:** The response correctly quoted both figures and added a valid computed comparison (\$590 million - \$538 million = \$52 million, approximately 9.7%); the validator required numbers to appear verbatim in the evidence and could not recognize the derived value.
- **Verifier verdict:** false positive (confidence: high)
- **Validator flags:** Unsupported number: '\$52' does not appear in the evidence

Failure 200 — `SEC\_10K\_multi\_neg\_1554\_01` · — confidence

DETERMINISTIC:ANSWERED UNANSWERABLE

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What was the net operating revenue increase for the EMEA operating segment?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** The EMEA operating segment's net operating revenues increased **5% in 2025 versus 2024**.
- **Diagnostic reviewer diagnosis:** **Negative-answer specialist** model error
- **Diagnostic reviewer root cause:** [code:answered_unanswerable] GT is UNANSWERABLE but the response asserts a concrete answer
- **Verifier verdict:** — (confidence: —)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 201 — `SEC\_10K\_multi\_1299\_01` · HIGH confidence

BAD QUESTION

- **Category:** Precision
- **Question:** What is the total senior debt at December 31, 2025?

- ✓ **Ground truth:** 226,071
- ✗ **Model answered:** Total senior debt at December 31, 2025 was **\$347.5 billion**:
- Parent company: **\$234.5 billion**
- Subsidiaries: **\$113.0 billion**
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** bad test
- **Diagnostic reviewer root cause:** The model interpreted "total senior debt" as the aggregate of fixed- and variable-rate senior debt for both the parent and subsidiaries, while the keyed answer copied the \$226,071 million total from only the parent-company fixed-rate senior-debt row. The validator also failed to recognize the model's computed aggregate because \$347.5 billion is not printed verbatim.
- **Verifier verdict:** bad test (confidence: high)
- **Validator flags:** Unsupported number: '\$347.5' does not appear in the evidence

Failure 202 — `SEC\_10K\_multi\_0202\_00` · HIGH confidence

UNIT/FORMAT MISMATCH

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Precision
- **Question:** What is the income before income tax expense?
- ✓ **Ground truth:** 72,595
- ✗ **Model answered:** JPMorgan Chase's income before income tax expense was **\$72.595 billion in 2025** (2024: \$75.081 billion; 2023: \$61.612 billion).
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** model error
- **Diagnostic reviewer root cause:** [verifier] The model hallucinates years (2025, 2024, 2023) that are not present in the provided evidence, which makes its answer incorrect.
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Unsupported number: '\$72.595' does not appear in the evidence

Failure 203 — `SEC\_10K\_multi\_neg\_1265\_03` · — confidence

DETERMINISTIC:ANSWERED UNANSWERABLE

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What is the total amount of goodwill allocated to the Asset & Wealth Management segment in 2024 that was not allocated in 2023?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** Asset & Wealth Management goodwill decreased by **\$61 million**, from **\$8.582 billion** in 2023 to **\$8.521 billion** in 2024.
- **Diagnostic reviewer diagnosis:** **Negative-answer specialist** model error
- **Diagnostic reviewer root cause:** [code:answered_unanswerable] GT is UNANSWERABLE but the response asserts a concrete answer
- **Verifier verdict:** — (confidence: —)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer; Unsupported number: '\$61' does not appear in the evidence

Failure 204 — `SEC\_10K\_multi\_1451\_01` · HIGH confidence

VALIDATOR OVER-FIRED

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream

- **Category:** Precision
- **Question:** What is the income from dividends from subsidiaries and affiliates in 2025?
- ✓ **Ground truth:** 50,000
- ✗ **Model answered:** Income from dividends from subsidiaries and affiliates in 2025 was **\$50.0 billion**.
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** false positive
- **Diagnostic reviewer root cause:** Validator failed to account for the table's 'in millions' units: 50,000 million dollars is exactly \$50.0 billion.
- **Verifier verdict:** false positive (confidence: high)
- **Validator flags:** Unsupported number: '\$50.0' does not appear in the evidence

Failure 205 — `SEC\_10K\_multi\_1159\_01` · — confidence

DETERMINISTIC:GT NUMBER CONFLICT

- **Severity:** HIGH — an incorrect quantitative result
- **Category:** Precision
- **Question:** What is the percentage of criticized to total retained loans secured by real estate for 2025?
- ✓ **Ground truth:** 4.43
- ✗ **Model answered:** The percentage was **6.37%** in 2025.
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** model error
- **Diagnostic reviewer root cause:** [code:gt_number_conflict] Short numeric GT; response asserts a different quantity
- **Verifier verdict:** — (confidence: —)
- **Validator flags:** (none)

Failure 206 — `SEC\_10K\_multi\_0852\_00` · — confidence

DETERMINISTIC:GT NUMBER CONFLICT

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Precision
- **Question:** What is the total amount of level 3 assets measured at fair value on a nonrecurring basis as of December 31, 2025?
- ✓ **Ground truth:** 863 million
- ✗ **Model answered:** The total amount was **\$1.392 billion** as of December 31, 2025.
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** model error
- **Diagnostic reviewer root cause:** [code:gt_number_conflict] Short numeric GT; response asserts a different quantity
- **Verifier verdict:** — (confidence: —)
- **Validator flags:** Unsupported number: '\$1.392' does not appear in the evidence

Failure 207 — `SEC\_10K\_multi\_1497\_00` · HIGH confidence

BAD QUESTION

- **Category:** Cross-Reference
- **Question:** What was the average interest rate earned on loans in U.S. operations for the year ended December 31, 2025, and what was the average balance of loans in U.S. operations for the same year?
- ✓ **Ground truth:** 6.86% and \$1,271,393
- ✗ **Model answered:** For 2025, the average interest rate earned on loans in U.S. operations was **6.86%**, and the average loan balance was **\$1,271,393 million** (approximately **\$1.27 trillion**).
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** bad test

- **Diagnostic reviewer root cause:** [verifier] The ground truth is incorrect because it omits the unit 'million' for the average balance; the evidence table header clearly states figures are '(in millions, except rates)', meaning the correct balance is \$1,271,393 million, which the model correctly identified.
- **Verifier verdict:** bad test (confidence: high)
- **Validator flags:** Unsupported number: '\$1.27' does not appear in the evidence

Failure 208 — `SEC\_10K\_multi\_neg\_0646\_02` · — confidence

DETERMINISTIC:ANSWERED UNANSWERABLE

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What is the allowance for loan losses to retained nonaccrual loans for the highest credit ratio?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** The highest allowance-for-loan-losses-to-retained-nonaccrual-loans ratio was **437% in 2023**.
- **Diagnostic reviewer diagnosis:** **Negative-answer specialist** model error
- **Diagnostic reviewer root cause:** [code:answered_unanswerable] GT is UNANSWERABLE but the response asserts a concrete answer
- **Verifier verdict:** — (confidence: —)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 209 — `SEC\_10K\_multi\_0910\_00` · HIGH confidence

BAD QUESTION

- **Category:** Cross-Reference
- **Question:** What is the total amount for Asset Managers and Consumer & Retail combined?
- ✓ **Ground truth:** 302,693
- ✗ **Model answered:** Asset Managers and Consumer & Retail combined total **\$286.793 billion** (\$152.848 billion + \$133.945 billion).
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** bad test
- **Diagnostic reviewer root cause:** The model summed the first listed total for Asset Managers (152,848) and Consumer & Retail (133,945), correctly obtaining 286,793; the validator rejected the derived sum because it does not appear verbatim in the evidence.
- **Verifier verdict:** bad test (confidence: high)
- **Validator flags:** Unsupported number: '\$286.793' does not appear in the evidence

Failure 210 — `SEC\_10K\_multi\_neg\_0728\_01` · — confidence

DETERMINISTIC:ANSWERED UNANSWERABLE

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** How much is the funding adjustment (FVA) for a typical credit derivative?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** No fixed "typical" FVA is provided. FVA depends on expected funding requirements, collateral arrangements, market funding costs, and the firm's credit risk. Aggregate derivatives FVA was a **\$18 million loss in 2025**, versus gains of **\$99 million in 2024** and **\$114 million in 2023**.
- **Diagnostic reviewer diagnosis:** **Negative-answer specialist** model error
- **Diagnostic reviewer root cause:** [code:answered_unanswerable] GT is UNANSWERABLE but the response asserts a concrete answer

- **Verifier verdict:** — (confidence: —)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 211 — `SEC\_10K\_multi\_1584\_01` · HIGH confidence

UNIT/FORMAT MISMATCH

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Cross-Reference
- **Question:** What is the total amount of equity attributable to shareowners of The Coca-Cola Company at the end of 2025, considering the equity attributable to shareowners at the beginning of the year and the net income attributable to shareowners of The Coca-Cola Company?
- ✓ **Ground truth:** 32,169
- ✗ **Model answered:** Equity attributable to Coca-Cola shareowners was **\$32.169 billion at the end of 2025**. It began at **\$24.856 billion**, and **\$13.107 billion** of net income was added, with dividends and other equity adjustments reducing the final balance.
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** false positive
- **Diagnostic reviewer root cause:** Validator failed to normalize units: the evidence reports 32,169 in millions, which is exactly \$32.169 billion as stated by the model.
- **Verifier verdict:** false positive (confidence: high)
- **Validator flags:** Unsupported number: '\$32.169' does not appear in the evidence

Failure 212 — `SEC\_10K\_multi\_neg\_1614\_04` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What is the name of the company's auditor?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** Ernst & Young LLP.
- **Diagnostic reviewer diagnosis:** **Negative-answer specialist** model error
- **Diagnostic reviewer root cause:** The model relied on likely prior knowledge or an assumption about Coca-Cola's auditor rather than restricting its answer to the provided evidence, which does not identify an auditor.
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 213 — `SEC\_10K\_multi\_1549\_00` · HIGH confidence

VALIDATOR OVER-FIRED

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Cross-Reference
- **Question:** What is the estimated increase in net periodic pension cost if the discount rate decreases by 50 basis points, and how does this compare to the estimated increase if the expected long-term rate of return on plan assets decreases by 50 basis points?
- ✓ **Ground truth:** \$9 million and \$19 million, respectively
- ✗ **Model answered:** A 50-basis-point decrease in the discount rate would increase 2026 net periodic pension cost by an estimated **\$9 million**.

A 50-basis-point decrease in the expected long-term return on plan assets would increase it by **\$19 million-\$10 million more**, or about **2.1 times** the discount-rate impact.

- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** false positive
- **Diagnostic reviewer root cause:** Validator rejected the correctly derived comparison of approximately 2.1 times because that number was calculated from the stated \$19 million and \$9 million figures rather than appearing verbatim in the evidence.
- **Verifier verdict:** false positive (confidence: high)
- **Validator flags:** Unsupported number: '2.1' does not appear in the evidence

Failure 214 — `SEC\_10K\_multi\_0253\_00` · — confidence

DETERMINISTIC:GT NUMBER CONFLICT

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Precision
- **Question:** What is the percentage change in total liabilities from 2024 to 2025?
- ✓ **Ground truth:** 11%
- ✗ **Model answered:** Total liabilities increased by approximately **8.4%** from 2024 to 2025:

$$\frac{(3,834,359 - 3,537,567)}{3,537,567} \times 100 \approx 8.4\%$$

- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** model error
- **Diagnostic reviewer root cause:** [code:gt_number_conflict] Short numeric GT; response asserts a different quantity
- **Verifier verdict:** — (confidence: —)
- **Validator flags:** Unsupported number: '8.4%' does not appear in the evidence

Failure 215 — `SEC\_10K\_multi\_neg\_1160\_00` · — confidence

DETERMINISTIC:ANSWERED UNANSWERABLE

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What is the total number of loans that are 90 or more days past due and still accruing?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** As of December 31, 2025, loans 90 or more days past due and still accruing totaled **\$2.732 billion** in the retained credit card portfolio.
- **Diagnostic reviewer diagnosis:** **Negative-answer specialist** model error
- **Diagnostic reviewer root cause:** [code:answered_unanswerable] GT is UNANSWERABLE but the response asserts a concrete answer
- **Verifier verdict:** — (confidence: —)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer; Unsupported number: '\$2.732' does not appear in the evidence

Failure 216 — `SEC\_10K\_multi\_1073\_00` · HIGH confidence

UNSUPPORTED NUMBER

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Precision
- **Question:** What is the total held-to-maturity securities value?

- ✓ **Ground truth:** 270,134
- ✗ **Model answered:** The total held-to-maturity securities value is **\$270.199 billion** in amortized cost (**\$253.344 billion** fair value).
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** model error
- **Diagnostic reviewer root cause:** Model made an arithmetic or transcription error, reporting \$270.199 billion instead of quoting the total row's \$270.134 billion amortized cost; the fair value figure was correct.
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Unsupported number: '\$270.199' does not appear in the evidence

Failure 217 — `SEC\_10K\_multi\_neg\_1554\_03` · — confidence

DETERMINISTIC:ANSWERED UNANSWERABLE

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What was the total net operating revenue increase for the Bottling Investments operating segment?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** Bottling Investments' net operating revenues **decreased 8%** in 2025 versus 2024.
- **Diagnostic reviewer diagnosis:** **Negative-answer specialist** model error
- **Diagnostic reviewer root cause:** [code:answered_unanswerable] GT is UNANSWERABLE but the response asserts a concrete answer
- **Verifier verdict:** — (confidence: —)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 218 — `SEC\_10K\_multi\_neg\_1169\_04` · — confidence

DETERMINISTIC:ANSWERED UNANSWERABLE

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What is the total amount of payment deferrals provided in 2023?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** The total amount of payment deferrals provided in 2023 was **\$690 million**:
- Secured by real estate: \$8 million
- Commercial and industrial: \$437 million
- Other loans: \$245 million
- **Diagnostic reviewer diagnosis:** **Negative-answer specialist** model error
- **Diagnostic reviewer root cause:** [code:answered_unanswerable] GT is UNANSWERABLE but the response asserts a concrete answer
- **Verifier verdict:** — (confidence: —)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer; Unsupported number: '\$8' does not appear in the evidence

Failure 219 — `SEC\_10K\_multi\_0570\_00` · HIGH confidence

UNSUPPORTED NUMBER

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Conditional

- **Question:** If the loan maturities are categorized by class, and the residential real estate loan class is considered, what is the estimated principal repayment amount for this class?
- ✓ **Ground truth:** calculated by weighted-average loan balance and interest rates for loan pools based on remaining loan term
- ✗ **Model answered:** The estimated principal repayment amount for residential real estate loans was **\$333.148 billion** as of December 31, 2025.
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** model error
- **Diagnostic reviewer root cause:** [verifier] The model hallucinated a specific monetary value (\$333.148 billion) that is not present in the evidence, which is a clear model error.
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Unsupported number: '\$333.148' does not appear in the evidence

Failure 220 — `SEC\_10K\_multi\_neg\_0039\_03` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What is the name of the CEO of NVIDIA Corporation?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** Jen-Hsun Huang.
- **Diagnostic reviewer diagnosis:** **Negative-answer specialist** model error
- **Diagnostic reviewer root cause:** The model relied on external knowledge about NVIDIA's leadership rather than restricting its answer to the provided evidence, which does not identify the CEO.
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 221 — `SEC\_10K\_multi\_0592\_01` · HIGH confidence

META LEAK

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Precision
- **Question:** What is the threshold for Metals & Mining?
- ✓ **Ground truth:** 406
- ✗ **Model answered:** No specific "threshold" is stated for Metals & Mining. Its wholesale credit exposure was **\$17.767 billion in 2025**, compared with **\$15.860 billion in 2024**.
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** model error
- **Diagnostic reviewer root cause:** The model read the Metals & Mining row's total exposure value (17,767) instead of the threshold column value (406), then added an unsupported year-over-year comparison.
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Unsupported number: '\$17.767' does not appear in the evidence

Failure 222 — `SEC\_10K\_multi\_0581\_00` · HIGH confidence

VALIDATOR OVER-FIRED

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Cross-Reference
- **Question:** What percentage of total loans is comprised of investment-grade loans, given the total amount of investment-grade loans and the total amount of all loans?

- ✓ **Ground truth:** 68%
- ✗ **Model answered:** Investment-grade loans comprise **68.32%** of total loans:

$$\frac{\$541.364 \text{ billion}}{\$792.367 \text{ billion}} \times 100 = 68.32\%$$

- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** false positive
- **Diagnostic reviewer root cause:** pre-Doctor equivalence: numeric match: model has 68.32% ~ GT 68%
- **Verifier verdict:** false positive (confidence: high)
- **Validator flags:** Unsupported number: '\$541.364' does not appear in the evidence

Failure 223 — `SEC\_10K\_multi\_0861\_00` · HIGH confidence

UNSUPPORTED NUMBER

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Precision
- **Question:** What is the value of Federal funds purchased and securities loaned or sold under repurchase agreements?
- ✓ **Ground truth:** 82.2
- ✗ **Model answered:** Federal funds purchased and securities loaned or sold under repurchase agreements were **\$442.396 billion as of December 31, 2025**, compared with **\$296.835 billion in 2024**.
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** model error
- **Diagnostic reviewer root cause:** The model retrieved figures from the wrong disclosure or table row/year instead of reading the requested financial-liabilities row, which reports 82.2.
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Unsupported number: '\$442.396' does not appear in the evidence

Failure 224 — `SEC\_10K\_multi\_neg\_0516\_00` · HIGH confidence

BAD QUESTION

- **Category:** Negative (hallucination traps)
- **Question:** What is the average interest rate on non-interest-bearing deposits at JPMorgan Chase's U.S. offices as of December 31, 2025?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** Not applicable (N/A). Noninterest-bearing deposits do not earn interest.
- **Diagnostic reviewer diagnosis:** **Negative-answer specialist** bad test
- **Diagnostic reviewer root cause:** Validator failed to treat the table's "NA" and the model's "Not applicable" as equivalent to an unanswerable/nonexistent average interest rate for noninterest-bearing deposits.
- **Verifier verdict:** bad test (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 225 — `SEC\_10K\_multi\_0934\_01` · HIGH confidence

TABLE MISREAD

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Precision
- **Question:** What is the total fair value of trading assets and liabilities before rounding?
- ✓ **Ground truth:** 663,566
- ✗ **Model answered:** The total fair value was **\$1,018,892 million (\$1.018892 trillion)**, comprising:

- Trading assets: **\$802,873 million**
- Trading liabilities: **\$216,019 million**
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** model error
- **Diagnostic reviewer root cause:** The model used trading asset and liability totals from the wrong table or reporting basis and summed them, rather than reading the applicable total fair value figure of 663,566 from the evidence row.
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Unsupported number: '\$1,018,892' does not appear in the evidence

Failure 226 — `SEC\_10K\_multi\_0942\_01` · HIGH confidence

BAD QUESTION

- **Category:** Conditional
- **Question:** If the derivative payables are U.S. GAAP nettable and interest rate contracts, what is the net amount?
- ✓ **Ground truth:** 7,840
- ✗ **Model answered:** The net derivative payables for U.S. GAAP-nettable interest rate contracts were **\$6.729 billion as of December 31, 2025.**
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** bad test
- **Diagnostic reviewer root cause:** The model used or aggregated the 2025 net amounts for interest-rate contracts, while the expected answer selects the 2024 OTC net derivative payable of \$7,840 million.
- **Verifier verdict:** bad test (confidence: high)
- **Validator flags:** Unsupported number: '\$6.729' does not appear in the evidence

Failure 227 — `SEC\_10K\_multi\_0467\_00` · HIGH confidence

VALIDATOR OVER-FIRED

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Cross-Reference
- **Question:** What is the total stockholders' equity as of December 31, 2025, and what is the common stockholders' equity as of the same date?
- ✓ **Ground truth:** The total stockholders' equity is \$362,438, and the common stockholders' equity is \$342,393.
- ✗ **Model answered:** As of December 31, 2025:
- **Total stockholders' equity:** **\$362.438 billion**
- **Common stockholders' equity:** **\$342.393 billion**
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** false positive
- **Diagnostic reviewer root cause:** Validator failed to recognize a valid unit conversion: the table reports amounts in millions, so \$362,438 million equals \$362.438 billion, and \$342,393 million equals \$342.393 billion.
- **Verifier verdict:** false positive (confidence: high)
- **Validator flags:** Unsupported number: '\$362.438' does not appear in the evidence

Failure 228 — `SEC\_10K\_multi\_0744\_00` · HIGH confidence

BAD QUESTION

- **Category:** Precision
- **Question:** What is the total derivative payables?
- ✓ **Ground truth:** 14,806

- **Model answered:** Total derivative payables were **\$596.738 billion** as of December 31, 2025, or **\$46.329 billion after netting adjustments**.
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** bad test
- **Diagnostic reviewer root cause:** The model summed the three fair-value hierarchy amounts for the explicit 'Total derivative payables' row (3,399 + 581,336 + 12,003 = 596,738) and also reported the row's net amount of 46,329; the validator failed to recognize the computed gross total because it was not printed verbatim.
- **Verifier verdict:** bad test (confidence: high)
- **Validator flags:** Unsupported number: '\$596.738' does not appear in the evidence

Failure 229 — `SEC\_10K\_multi\_0825\_00` · HIGH confidence

TABLE MISREAD

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Precision
- **Question:** What is the total number of mortgage-backed securities?
- **Ground truth:** 771
- **Model answered:** As of December 31, 2025, total mortgage-backed securities classified as trading assets were **\$162.085 billion**.
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** model error
- **Diagnostic reviewer root cause:** Model retrieved an unrelated mortgage-backed-securities figure, likely from the wrong table or reporting period, instead of reading the 'Total mortgage-backed securities' row value of 771 in the provided evidence.
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Unsupported number: '\$162.085' does not appear in the evidence

Failure 230 — `SEC\_10K\_multi\_1173\_00` · HIGH confidence

BAD QUESTION

- **Category:** Cross-Reference
- **Question:** What is the percentage of loan modifications to total retained Commercial and industrial loans, and how does it compare to the number of single modifications?
- **Ground truth:** 0.71% of loan modifications to total retained Commercial and industrial loans, with 1,180 single modifications.
- **Model answered:** For 2025, Commercial and industrial loan modifications represented **1.42%** of total retained C&I loans. Clearly identified single modifications totaled **\$1.997 billion**-\$1.308 billion of term extensions plus \$689 million of payment deferrals-representing **1.15%** of retained C&I loans and about **81%** of the total \$2.461 billion modified balance.
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** bad test
- **Diagnostic reviewer root cause:** The model used figures from the wrong reporting period/table and aggregated modification balances instead of extracting the 2024 single-modification term-extension row showing \$1,180 million and 0.71%.
- **Verifier verdict:** bad test (confidence: high)
- **Validator flags:** Unsupported number: '1.42%' does not appear in the evidence

Failure 231 — `SEC\_10K\_multi\_0724\_01` · HIGH confidence

UNSUPPORTED NUMBER

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Precision

- **Question:** What is the total liabilities and stockholders' equity?
- ✓ **Ground truth:** \$4,424,900
- ✗ **Model answered:** Total liabilities and stockholders' equity were **\$4,424,900 million** (approximately **\$4.42 5 trillion**) as of December 31, 2025.
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** model error
- **Diagnostic reviewer root cause:** [verifier] The model hallucinates the unit 'million' and a date ('December 31, 2025'), which are not present in the evidence and fundamentally change the meaning of the answer.
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Unsupported number: '\$4.425' does not appear in the evidence

Failure 232 — `SEC\_10K\_multi\_0630\_01` · HIGH confidence

BAD QUESTION

- **Category:** Precision
- **Question:** What is the total for loans due after one year at fixed interest rates?
- ✓ **Ground truth:** 14,356
- ✗ **Model answered:** The total consumer loans due after one year at fixed interest rates was **\$192.066 billion**.
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** bad test
- **Diagnostic reviewer root cause:** The model read or aggregated an unrelated consumer-loan figure instead of extracting the relevant fixed-rate maturity value from the displayed wholesale-loan table.
- **Verifier verdict:** bad test (confidence: high)
- **Validator flags:** Unsupported number: '\$192.066' does not appear in the evidence

Failure 233 — `SEC\_10K\_multi\_neg\_0080\_02` · HIGH confidence

HALLUCINATION

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What is the name of the regulatory agency conducting the inquiries and investigations mentioned in the passage?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** The U.S. Securities and Exchange Commission (SEC).
- **Diagnostic reviewer diagnosis:** **Negative-answer specialist** model error
- **Diagnostic reviewer root cause:** The model inferred a plausible regulator—the SEC—from the securities-law context and references to Exchange Act violations, even though the passage does not identify which agency conducts the regulatory and government inquiries and investigations.
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 234 — `SEC\_10K\_multi\_1158\_01` · HIGH confidence

MISINTERPRETED

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Cross-Reference
- **Question:** What is the total amount of gross charge-offs for the year 2024, and how does it compare to the total gross charge-offs for the year 2023?
- ✓ **Ground truth:** The total gross charge-offs for 2024 is \$-, and for 2023 is \$38.

- **Model answered:** Gross charge-offs were **\$43 million for 2024-originated loans**, compared with **\$64 million for 2023-originated loans**-a decrease of **\$21 million**, or approximately **32.8%**.
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** model error
- **Diagnostic reviewer root cause:** The model misread the origination-year columns and supplied unsupported figures, then calculated a percentage decrease from those incorrect values. The gross charge-offs row shows no amount for 2024 and \$38 million for 2023.
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Unsupported number: '32.8%' does not appear in the evidence

Failure 235 — `SEC\_10K\_multi\_0259\_01` · HIGH confidence

UNSUPPORTED NUMBER

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Precision
- **Question:** What is the income tax expense?
- **Ground truth:** 15,547
- **Model answered:** JPMorgan Chase's income tax expense was **\$15.547 billion for 2025**.
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** model error
- **Diagnostic reviewer root cause:** [verifier] The model's answer fabricates information not present in the evidence, specifically the year "2025" and the company name "JPMorgan Chase's".
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Unsupported number: '\$15.547' does not appear in the evidence

Failure 236 — `SEC\_10K\_multi\_neg\_0618\_04` · — confidence

DETERMINISTIC:ANSWERED UNANSWERABLE

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What is the dollar amount of net portfolio activity as of December 31, 2025?
- **Ground truth:** UNANSWERABLE
- **Model answered:** The net portfolio activity was **\$747 million** as of December 31, 2025.
- **Diagnostic reviewer diagnosis:** **Negative-answer specialist** model error
- **Diagnostic reviewer root cause:** [code:answered_unanswerable] GT is UNANSWERABLE but the response asserts a concrete answer
- **Verifier verdict:** — (confidence: —)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 237 — `SEC\_10K\_multi\_1436\_00` · HIGH confidence

BAD QUESTION

- **Category:** Precision
- **Question:** What is the total noninterest expense?
- **Ground truth:** 38,036
- **Model answered:** Total noninterest expense was **\$95.64 billion in 2025**.
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** bad test
- **Diagnostic reviewer root cause:** The model selected or aggregated a 2025 company-wide expense figure instead of extracting the intended table cell, 38,036.
- **Verifier verdict:** bad test (confidence: high)
- **Validator flags:** Unsupported number: '\$95.64' does not appear in the evidence

Failure 238 — `SEC\_10K\_multi\_1335\_01` · HIGH confidence

BAD QUESTION

- **Category:** Calculation
- **Question:** What is the aggregate purchase price of common stock repurchases in 2023?
- ✓ **Ground truth:** 69.5
- ✗ **Model answered:** 9898
- **Diagnostic reviewer diagnosis:** **Calculation specialist** bad test
- **Diagnostic reviewer root cause:** Ground truth read 69.5 from the preceding row for the number of shares repurchased, rather than the aggregate purchase price row; 9898 is likely the purchase price in millions of dollars.
- **Verifier verdict:** bad test (confidence: high)
- **Validator flags:** Number mismatch: expected 69.5, not found in response

Failure 239 — `SEC\_10K\_multi\_0477\_00` · — confidence

DETERMINISTIC:GT NUMBER CONFLICT

- **Severity:** HIGH — an incorrect quantitative result
- **Category:** Precision
- **Question:** What was the amount of dividends declared on common stock for the year ended December 31, 2025?
- ✓ **Ground truth:** -16,060
- ✗ **Model answered:** The disclosed quarterly common stock dividend was **\$1.50 per share**, declared on December 9, 2025.
- **Diagnostic reviewer diagnosis:** **General diagnostic specialist** model error
- **Diagnostic reviewer root cause:** [code:gt_number_conflict] Short numeric GT; response asserts a different quantity
- **Verifier verdict:** — (confidence: —)
- **Validator flags:** (none)

Failure 240 — `SEC\_10K\_multi\_neg\_1409\_04` · — confidence

DETERMINISTIC:ANSWERED UNANSWERABLE

- **Severity:** HIGH — a confident, specific but unsupported answer — the costliest class: it reads as authoritative and is hard to catch downstream
- **Category:** Negative (hallucination traps)
- **Question:** What is the percentage of risk participations for standby letters of credit compared to other letters of credit?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** In 2025, standby letters of credit had **approximately 1,790%** of the risk participations of other letters of credit:

$$\frac{\$9.2 \times 10^9}{\$514 \times 10^6} \times 100 \approx 1,790\%$$

That is about **17.9 times** as much.

- **Diagnostic reviewer diagnosis:** **Negative-answer specialist** model error
- **Diagnostic reviewer root cause:** [code:answered_unanswerable] GT is UNANSWERABLE but the response asserts a concrete answer
- **Verifier verdict:** — (confidence: —)

- **Validator flags:** Hallucination: model answered a question the evidence cannot answer; Unsupported number: '1,790%' does not appear in the evidence

Failure 241 — `SEC\_10K\_multi\_0275\_01` · HIGH confidence

BAD QUESTION

- **Category:** Calculation
- **Question:** What is TBVPS for December 31, 2025?
- ✓ **Ground truth:** 0.845
- ✗ **Model answered:** 107.56
- **Diagnostic reviewer diagnosis:** **Calculation specialist** bad test
- **Diagnostic reviewer root cause:** The model appears to have used the TBVPS per-share figure from the reconciliation, while the benchmark mapped an unrelated value (0.845) to the question.
- **Verifier verdict:** bad test (confidence: high)
- **Validator flags:** Number mismatch: expected 0.845, not found in response

