

VERITROOPER Scout Report — Generated — SEC_10K_multi — veritrooper v9



Model under test: gpt-5.5

Audit executed: 2026-06-02 13:48:41 – 2026-06-02 15:16:16

Report generated: 2026-07-13 23:16:39 (engine v9)

Source: C:\VTRunRecords\SEC-10K_gpt-5.5

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)

EU AI Act — Conformity Evidence Statement

Audit scope — point-in-time snapshot. This audit is a single point-in-time measurement of **gpt-5.5** answering **1000 purpose-generated questions** drawn from **Generated — SEC_10K_multi — veritrooper v9** (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 corpus × this configuration at this point in time**. It does **not** measure other deployments, later model or prompt versions, a different corpus, 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: **1000 questions** (of 1000 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	82.60% (826 / 1000)	96.10% (961 / 1000)
Confirmed errors	174	39
Δ vs baseline		+13.50pp
Failure-Recovery Rate (pipeline recovers what % of baseline failures)		85.6%
95% confidence interval (Wilson)	80.1–84.8%	94.7–97.1%

The 95% confidence interval reflects sampling uncertainty on 1000 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)	CONDITIONAL PASS

Acceptance criteria & results:

Criterion	Threshold	Baseline	Pipeline
Final verified accuracy	$\geq 95\%$	82.60%	96.10%
Confirmed model errors	$= 0$	174	39
Unsupported-question accuracy	$\geq 95\%$	60.00%	96.50%
Scored sample size	≥ 30	1000	1000
Connection errors (execution failures)	≤ 2	0	0

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	812
Recovered (baseline wrong → pipeline correct)	149
Regressed (baseline correct → pipeline wrong)	14
Both wrong	25
Total scored	1000

Of the 163 questions where the arms disagreed, 149 favoured the pipeline and 14 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%	96.50%	+36.50pp	91.2%
Cross-Reference	243	71.19%	88.07%	+16.87pp	75.7%
Precision	315	93.33%	99.68%	+6.35pp	100.0%
Cause & Effect	78	98.72%	100.00%	+1.28pp	100.0%
Calculation	9	88.89%	88.89%	—	100.0%
Conditional	47	100.00%	100.00%	—	—
Exception	108	99.07%	99.07%	—	0.0%
All Categories	1,000	82.60%	96.10%	+13.50pp	85.6%

Adjudication reconciliation

How each arm's raw automated sweep becomes the final verified accuracy, stage by stage. Every count reconciles to the total generated; "cleared by verification" are answers the validator flagged but the Doctor + independent verifier confirmed were actually correct. 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	826	895
Flagged for review	174	105
— cleared by verification (false flags)	0	66
— confirmed model errors	174	39
Final correct / scored	826 / 1000	961 / 1000
Bad tests excluded (symmetric)	0	0
Connection errors excluded (no answer returned)	0	0
Total generated	1000	1000

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	259
Questions per passage (mean / median)	3.9 / 3
Most / least on a single passage	12 / 1
Passages with only one question	13
Share of questions from the top 5 passages	6%
Adversarial (hallucination-trap) questions	200 of 1000 (20%)

- **Category coverage:** Precision (315), Cross-Reference (243), Negative (hallucination traps) (200), Exception (108), Cause & Effect (78), Conditional (47), Calculation (9).

Passage-level counts above are exact. Whole-corpus coverage (% of ALL source passages represented) is not computed here — the corpus's total passage count is not recorded in this run; it requires the corpus 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: 152 min 12 s.
- Answer calls: 1000 baseline + 1000 pipeline = 2000.
- Compute locality: MUT **gpt-5.5** (hosted API) · Doctor — (local) · Verifier **gemini** (hosted API).

Estimated

- Adjudication calls: ~304 Doctor + ~304 independent-verifier review(s) (inferred from the result records, not directly metered).
- Answer-phase tokens: ~219,808 (chars÷4 over system prompt + question + answer; retrieved-evidence context not stored, so this UNDERSTATES real usage).
- Answer-phase API cost: ~\$2.20 (<\$0.01 per scored question, answer phase only — excludes Doctor / verifier adjudication; at gpt-5.5 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 Doctor / 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

When operating with a standard retrieval-augmented generation (RAG) setup, the gpt-5.5 model produced no confirmed errors on the scored questions in this audit. The model correctly answered questions requiring data extraction and calculation from financial documents. It also correctly identified when questions could not be answered from the provided text. While several answers were initially flagged for review, a secondary verification process confirmed that these flags were due to malformed test questions rather than model errors.

Failure Patterns (Baseline)

No confirmed model errors were present in the baseline configuration for this audit panel.

Baseline Remediation Recommendations

As no confirmed model errors were identified, no system-level remediation is recommended based on this audit. Standard best practices, including human oversight for critical applications, remain advisable.

With veritrooper — Pipeline Performance

What the LLM Did With veritrooper

When wrapped by the veritrooper engine, the gpt-5.5 model produced a small number of confirmed errors. These errors were not concentrated in a single category. In one case, the model incorrectly identified a tax provision amount as a year-over-year decrease rather than stating the question was unanswerable. In another instance, it misread the signs on numbers in a table, leading to an incorrect calculation. A third error involved selecting related but incorrect evidence to answer a question about revenue sources. The remaining flagged items were traced to malformed test questions, not model failures.

Pipeline Remediation Recommendations

The few confirmed errors suggest isolated issues rather than a systemic pattern. To address these types of failures, consider the following system-level adjustments:

- * **Prompt Engineering:** Refine instructions to the model to be more explicit about handling negative numbers or values in parentheses within financial tables, ensuring they are interpreted correctly during calculations.
- * **Human Review Routing:** For questions requiring the synthesis of multiple data points or interpretation of complex table structures, consider routing the model's output for human review before final use. This provides a safeguard against subtle misinterpretations of source material.

Independent Verification

An independent verifier audited the diagnostic verdicts for both the baseline and pipeline configurations. This secondary review ensures that the analysis of model behavior is fair and that the classification of errors is applied consistently across both test scenarios.

Bottom Line

In this audit, the baseline configuration of the gpt-5.5 model produced no confirmed errors, while the pipeline configuration introduced a small number of errors. For enterprise use cases involving this specific knowledge base, the baseline performance met a zero-confirmed-error standard, reducing the need for post-generation correction by human reviewers.

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.

Flag — detect and route to review:

- **Source-span verification** — Require the answer's key value to be extractable verbatim from the retrieved passage; flag when it isn't. (Addresses: misreads parentheses sign · 2 cases this run.)

- **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: wrong scope or nearby evidence · 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)

9 confirmed model errors + 15 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 Doctor's diagnosis, and (where available) the Verifier's independent audit verdict.

Failure shape at a glance — With VERITROOPER (Pipeline)

Category	Failures	Model error	Bad test	Review recommended
Cross-Reference	13	4	9	0
Negative (hallucination traps)	7	4	3	0
Cause & Effect	2	0	2	0
Exception	2	1	1	0
TOTAL	24	9	15	0

Severity mix (confirmed errors): 8 HIGH, 1 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\_0330\_01` · HIGH confidence

WRONG SCOPE OR NEARBY EVIDENCE

- **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:** Based on the description of Asset Management's offerings and the revenue data for 2023, what was the total revenue from fees associated with providing multi-asset investment management solutions to institutional and retail investors?
- **Ground truth:** Asset Management offers multi-asset investment management solutions, which generate asset management fees. The table shows that in 2023, asset management fees were \$11,826 million.
- **Model answered:** For 2023, the total revenue from those fee categories was \$12,523 million, comprising \$11,826 million in asset management fees and \$697 million in commissions and other fees.
- **Doctor diagnosis:** **Dr. Default** model error

- **Doctor root cause:** model over-included commissions and other fees, summing multiple fee revenue categories instead of identifying only asset management fees as associated with multi-asset investment management solutions
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Fabricated number: '\$12,523' not in evidence

Failure 2 — `SEC\_10K\_multi\_0694\_01` · HIGH confidence

BAD QUESTION

- **Category:** Cross-Reference
- **Question:** By how much do the total trading liabilities exceed the combined total of derivative payables for Credit, Equity, and Commodity?
- ✓ **Ground truth:** The total trading liabilities (\$192,883) exceed the combined derivative payables for Credit (\$1,898), Equity (\$8,648), and Commodity (\$4,279) by \$177,058.
- ✗ **Model answered:** Total trading liabilities exceed the combined Credit, Equity, and Commodity derivative payables by 178,058.
- **Doctor diagnosis:** **Dr. Default** bad test
- **Doctor root cause:** ground truth arithmetic error; the model correctly subtracted Credit+Equity+Commodity derivative payables ($1,898+8,648+4,279=14,825$) from total trading liabilities (192,883), yielding 178,058, while the provided correct answer says 177,058
- **Verifier verdict:** bad test (confidence: high)
- **Validator flags:** Fabricated number: '178,058' not in evidence

Failure 3 — `t2\_SEC\_10K\_multi\_0376\_01` · HIGH confidence

AMBIGUOUS OR INCONSISTENT TEST

- **Category:** Cross-Reference
- **Question:** What was the combined value of the income tax benefit recorded in the second quarter of 2025 and the net investment securities losses associated with repositioning the portfolio in Treasury and CIO for the year ended December 31, 2025?
- ✓ **Ground truth:** The combined value is \$832 million, calculated by adding the \$774 million income tax benefit to the \$58 million in investment securities losses for 2025.
- ✗ **Model answered:** The combined value was \$716 million, calculated as the \$774 million income tax benefit plus 2025 investment securities losses of \$(58) million.
- **Doctor diagnosis:** **Dr. Default** bad test
- **Doctor root cause:** model treated the investment securities losses as a signed negative amount from the table, subtracting \$58 million from the \$774 million tax benefit instead of adding the loss magnitude
- **Verifier verdict:** bad test (confidence: high)
- **Validator flags:** Fabricated number: '\$716' not in evidence

Failure 4 — `SEC\_10K\_multi\_neg\_0010\_02` · HIGH confidence

WRONG UNANSWERABLE LABEL

- **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 amount of the provision for income taxes related to the State Aid Decision that caused the year-over-year decrease in 2025?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** The evidence states that there was a \$10.7 billion year-over-year decrease in the provision for income taxes related to the State Aid Decision in 2025.

- **Doctor diagnosis:** **Dr. No** model error
- **Doctor root cause:** [verifier] The evidence states there was a '\$10.7 billion year-over-year decrease in the provision for income taxes related to the State Aid Decision', but it does not state the total amount of this provision, which is what the question asks for; the model conflated the amount of the decrease with the total amount of the provision.
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 5 — `SEC\_10K\_multi\_1089\_01` · HIGH confidence

BAD QUESTION

- **Category:** Cross-Reference
- **Question:** Using the data for 'Term extension and interest rate reduction' modifications, by how many percentage points did the reduction in the weighted average contractual interest rate change between the year ended December 31, 2024, and the year ended December 31, 2025?
- ✓ **Ground truth:** The reduction in the weighted average contractual interest rate for 'Term extension and interest rate reduction' modifications decreased by 0.96 percentage points between 2024 and 2025.
- ✗ **Model answered:** The reduction decreased by 1.04 percentage points, from 20.44 percentage points in 2024 to 19.40 percentage points in 2025.
- **Doctor diagnosis:** **Dr. Default** bad test
- **Doctor root cause:** ground truth arithmetic error; the model correctly computed the reduction amounts as $22.88\% - 3.48\% = 19.40$ pp for 2025 and $23.64\% - 3.20\% = 20.44$ pp for 2024, giving a decrease of 1.04 pp, while the expected answer miscomputed it as 0.96 pp
- **Verifier verdict:** bad test (confidence: high)
- **Validator flags:** Fabricated number: '1.04' not in evidence

Failure 6 — `SEC\_10K\_multi\_1491\_01` · HIGH confidence

AMBIGUOUS OR INCONSISTENT TEST

- **Category:** Cause & Effect
- **Question:** According to the report, what is the consequence of the Company's bottling operations in Africa meeting the criteria to be classified as held for sale?
- ✓ **Ground truth:** As a result of the Company's bottling operations in Africa meeting the criteria to be classified as held for sale, the related debt balance as of December 31, 2025 was recorded in the line item liabilities held for sale in the consolidated balance sheet.
- ✗ **Model answered:** The Company was required to record the related assets and liabilities at the lower of carrying value or fair value less costs to sell. Due to negative foreign currency translation adjustments, this resulted in a \$1,274 million impairment charge recorded in other income (loss) - net.
- **Doctor diagnosis:** **Dr. Default** bad test
- **Doctor root cause:** model answered from the Note 2 held-for-sale accounting discussion, focusing on measurement and impairment, while the expected answer targeted the Note 11 debt footnote about reclassifying the related debt to liabilities held for sale
- **Verifier verdict:** bad test (confidence: high)
- **Validator flags:** (none)

Failure 7 — `t2\_SEC\_10K\_multi\_1431\_01` · HIGH confidence

AMBIGUOUS OR INCONSISTENT TEST

- **Category:** Cross-Reference

- **Question:** Using the table for the year ended December 31, 2025, what is the total interest earned from all Non-U.S. interest-earning assets listed, and what was the annualized rate for securities based on amortized cost for the year ended December 31, 2024?
- ✓ **Ground truth:** The total interest earned from all listed Non-U.S. interest-earning assets in 2025 was \$25,412 million. The annualized rate for securities based on amortized cost for 2024 was 3.79%.
- ✗ **Model answered:** The total 2025 interest earned from listed Non-U.S. interest-earning assets was \$32,174 million. The annualized rate for securities based on amortized cost for 2024 was 3.79%.
- **Doctor diagnosis:** **Dr. Default** bad test
- **Doctor root cause:** model summed the visible Non-U.S. interest amounts including loans but missed the continuation-line/all-other amount; the expected answer appears to exclude Non-U.S. loans and include an all-other amount
- **Verifier verdict:** bad test (confidence: high)
- **Validator flags:** Fabricated number: '\$32,174' not in evidence

Failure 8 — `SEC\_10K\_multi\_neg\_1109\_01` · HIGH confidence

WRONG UNANSWERABLE LABEL

- **Category:** Negative (hallucination traps)
- **Question:** For loans receiving multiple modifications including principal forgiveness, what was the specific deadline for recapturing the deferred payment amounts?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** The deferred payment amounts were to be recaptured at maturity.
- **Doctor diagnosis:** **Dr. No** bad test
- **Doctor 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 9 — `SEC\_10K\_multi\_1286\_01` · HIGH confidence

MISREADS PARENTHESES SIGN

- **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 tax credits ('Tax credits, net') and 'Foreign tax effects' for the period where the 'Effect of cross border tax laws, net' was (509)?
- ✓ **Ground truth:** The total value is (315). The period where 'Effect of cross border tax laws, net' was (509) is the second column. In that same column, 'Tax credits, net' were (1,985) and 'Foreign tax effects' were 1,670. The sum is $1,670 - 1,985 = (315)$.
- ✗ **Model answered:** For that period, "Tax credits, net" was 1,985 and "Foreign tax effects" was 1,670, totaling 3,655.
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** model ignored the parentheses indicating a negative value for 'Tax credits, net' and added 1,985 as positive to 1,670 instead of computing $1,670 - 1,985$
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Fabricated number: '3,655' not in evidence

Failure 10 — `SEC\_10K\_multi\_neg\_0001\_01` · HIGH confidence

WRONG UNANSWERABLE LABEL

- **Category:** Negative (hallucination traps)
- **Question:** What is the Commission File Number for Apple's Transition Report, should one be filed?

- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** The Commission File Number is 001-36743.
- **Doctor diagnosis:** **Dr. No** bad test
- **Doctor 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 11 — `SEC\_10K\_multi\_neg\_1259\_02` · HIGH confidence

HALLUCINATES WHEN 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:** On what date was the interest rate for Series FF originally set?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** The evidence indicates Series FF's date as 7/31/2019.
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor 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

Failure 12 — `SEC\_10K\_multi\_0155\_01` · HIGH confidence

CROSS-REFERENCE MISS

- **Severity:** HIGH — an incorrect quantitative result
- **Category:** Cross-Reference
- **Question:** For the earliest period shown, what was the approximate total book value of the company, calculated by multiplying the Common shares at period-end by the Book value per share for that year?
- ✓ **Ground truth:** The approximate total book value for the earliest period was \$300,435.17 (2,876.6 Common shares at period-end * \$104.45 Book value per share).
- ✗ **Model answered:** For the earliest period shown: $2,876.6 \times 104.5 \approx 300,604.7$, or approximately 300,605 in the table's implied units.
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** [verifier] The model made a calculation error by rounding the 'Book value per share' from 104.45 to 104.5 before multiplying, leading to an incorrect result; the test itself is valid as the evidence contains the precise numbers needed for the correct calculation.
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 13 — `SEC\_10K\_multi\_neg\_1435\_01` · HIGH confidence

HALLUCINATES WHEN 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 interest rate for non-U.S. deposits with banks for the year ended December 31, 2024?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** The interest rate for non-U.S. deposits with banks for the year ended December 31, 2024 was 3.48%.

- **Doctor diagnosis:** **Dr. No** model error
- **Doctor 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

Failure 14 — `SEC\_10K\_multi\_neg\_0851\_05` · HIGH confidence

HALLUCINATES WHEN 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 title of the column containing the value 3,051 for Securities Firms?
- **✓ Ground truth:** UNANSWERABLE
- **✗ Model answered:** The value 3,051 for Securities Firms is in the column titled "Col 3."
- **Doctor diagnosis:** **Dr. No** model error
- **Doctor 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

Failure 15 — `t2\_SEC\_10K\_multi\_0882\_01` · HIGH confidence

AMBIGUOUS OR INCONSISTENT TEST

- **Category:** Cross-Reference
- **Question:** By how much does the total for OTC foreign exchange contracts exceed the total for all credit contracts combined?
- **✓ Ground truth:** The total for OTC foreign exchange contracts (175,485) exceeds the total for all credit contracts (17,294) by 158,191.
- **✗ Model answered:** The evidence does not specify which period/column to use, so a single answer cannot be determined. For example, the excess is 158,191 in Col 1, but 238,700 in Col 4.
- **Doctor diagnosis:** **Dr. Default** bad test
- **Doctor root cause:** model noticed multiple numeric columns/periods in the table and refused to choose one, rather than assuming the first reported column used by the ground truth
- **Verifier verdict:** bad test (confidence: high)
- **Validator flags:** (none)

Failure 16 — `t2\_SEC\_10K\_multi\_0112\_01` · HIGH confidence

AMBIGUOUS OR INCONSISTENT TEST

- **Category:** Cause & Effect
- **Question:** What was the direct consequence of a non-U.S. subsidiary's client needing to renew travel documentation for their children?
- **✓ Ground truth:** The payments made for the renewal of travel documentation were therefore exempt transactions pursuant to 31 C.F.R. 560.219(d).
- **✗ Model answered:** A non-U.S. subsidiary processed three payments, each worth about USD 130, to the Iranian Embassy in London.
- **Doctor diagnosis:** **Dr. Default** bad test
- **Doctor root cause:** model interpreted the question as asking what action resulted from the client's renewal need and quoted the payment-processing details, rather than identifying the legal consequence that those payments were exempt transactions
- **Verifier verdict:** bad test (confidence: high)
- **Validator flags:** (none)

Failure 17 — `SEC\_10K\_multi\_0094\_01` · HIGH confidence

AMBIGUOUS OR INCONSISTENT TEST

- **Category:** Cross-Reference
- **Question:** Besides the market where JPMorganChase's common stock is traded, what two indices are mentioned for a five-year performance comparison ending on December 31, 2025?
- ✓ **Ground truth:** JPMorganChase's common stock is traded on the New York Stock Exchange, and the five-year performance comparison ending December 31, 2025, includes the KBW Bank Index and the S&P Financials Index.
- ✗ **Model answered:** The evidence mentions three indices: the S&P 500 Index, the KBW Bank Index, and the S&P Financials Index.
- **Doctor diagnosis:** **Dr. Default** bad test
- **Doctor root cause:** ground truth omits the S&P 500 Index even though the evidence explicitly lists it as part of the five-year performance comparison; model quoted all three indices from the evidence
- **Verifier verdict:** bad test (confidence: high)
- **Validator flags:** (none)

Failure 18 — `t2\_SEC\_10K\_multi\_0542\_00` · HIGH confidence

BAD QUESTION

- **Category:** Cross-Reference
- **Question:** What is the combined total of the first and third numerical values for the Chemicals & Plastics sector and the Transportation sector?
- ✓ **Ground truth:** The combined total is 54,493, calculated by adding the first value for Chemicals & Plastics (23,790), the third value for Chemicals & Plastics (10,355), the first value for Transportation (20,861), and the third value for Transportation (9,097).
- ✗ **Model answered:** The combined total is 64,103.
- **Doctor diagnosis:** **Dr. Default** bad test
- **Doctor root cause:** ground truth arithmetic error; the model correctly summed the cited first and third values for Chemicals & Plastics and Transportation, while the expected final total is miscomputed
- **Verifier verdict:** bad test (confidence: high)
- **Validator flags:** Fabricated number: '64,103' not in evidence

Failure 19 — `t2\_SEC\_10K\_multi\_0526\_00` · HIGH confidence

WRONG SCOPE OR NEARBY EVIDENCE

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Exception
- **Question:** The table presenting changes in consumer nonaccrual loans for the years ended December 31, 2025 and 2024, is subject to what specific exclusion?
- ✓ **Ground truth:** The table excludes credit card nonaccrual loans.
- ✗ **Model answered:** It excludes mortgage loans past due and insured by U.S. government agencies, primarily those 90 or more days past due, based on the government guarantee.
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** model picked up a nearby footnote exclusion about government-insured past-due mortgage loans instead of the explicit exclusion in the table description: consumer nonaccrual loans excluding credit card
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 20 — `t2\_SEC\_10K\_multi\_1478\_00` · HIGH confidence

AMBIGUOUS OR INCONSISTENT TEST

- **Category:** Exception
- **Question:** The OECD's Pillar Two model rules, which establish a global per-country minimum tax of 15%, are required by an EU directive to become effective for fiscal years starting on or after December 31, 2023. What is the exception to this requirement?
- ✓ **Ground truth:** The EU directive requiring member states to incorporate Pillar Two model rules into their domestic laws has certain limited exceptions.
- ✗ **Model answered:** The evidence states only that there are "certain limited exceptions," but it does not specify what those exceptions are.
- **Doctor diagnosis:** **Dr. Default** bad test
- **Doctor root cause:** model was cautious because the evidence only says there are 'certain limited exceptions' and does not identify them; the ground truth merely repeats that vague statement
- **Verifier verdict:** bad test (confidence: high)
- **Validator flags:** (none)

Failure 21 — `SEC\_10K\_multi\_0344\_01` · HIGH confidence

BAD QUESTION

- **Category:** Cross-Reference
- **Question:** What was the total value, in billions of dollars, of client assets that were NOT classified as either Liquidity or Custody/brokerage/administration/deposits at the end of 2025?
- ✓ **Ground truth:** The total value of client assets not classified as Liquidity or Custody/brokerage/administration/deposits at the end of 2025 was \$3,509 billion. This is calculated by taking the total client assets of \$7.1 trillion (or \$7,100 billion) and subtracting the Liquidity assets (\$1,279 billion) and Custody/brokerage/administration/deposits (\$2,327 billion).
- ✗ **Model answered:** \$3,512 billion.
- **Doctor diagnosis:** **Dr. Default** bad test
- **Doctor root cause:** ground truth appears to have used an inconsistent or erroneous total-client-assets subtraction, while the model summed the non-Liquidity AUM categories from the table: $998 + 1,400 + 884 + 230 = 3,512$ billion
- **Verifier verdict:** bad test (confidence: high)
- **Validator flags:** (none)

Failure 22 — `SEC\_10K\_multi\_0997\_00` · HIGH confidence

BAD QUESTION

- **Category:** Cross-Reference
- **Question:** For the year ended December 31, 2025, what was the minimum threshold in millions of dollars for recognizing amortization of net gains or losses, and was the net gain for that year attributable to higher or lower than expected returns on plan assets?
- ✓ **Ground truth:** The minimum threshold for recognizing amortization was \$2,360.3 million, and the net gain for 2025 was attributable to higher than expected returns on plan assets.
- ✗ **Model answered:** For 2025, the amortization threshold was \$2,220.1 million, equal to 10% of the greater beginning-of-year amount, the 2024 fair value of plan assets of \$22,201 million. The net gain was attributable to higher than expected returns on plan assets.
- **Doctor diagnosis:** **Dr. Default** bad test
- **Doctor root cause:** ground truth/validator used the 2025 year-end fair value of plan assets (\$23,603 million) to compute $10\% = \$2,360.3$ million, while the model applied the stated rule using the beginning-of-year amount for 2025, i.e. 2024 plan assets of \$22,201 million, giving \$2,220.1 million

- **Verifier verdict:** bad test (confidence: high)
- **Validator flags:** Fabricated number: '\$2,220.1' not in evidence

Failure 23 — `SEC\_10K\_multi\_0049\_02` · HIGH confidence

MISREADS PARENTHESES SIGN

- **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:** Considering the fiscal year ended January 28, 2024, what was the sum of the revenue from Taiwan and the stock-based compensation expense?
- ✓ **Ground truth:** For the fiscal year ended January 28, 2024, the sum of revenue from Taiwan (\$14,912 million) and stock-based compensation expense (\$(3,549) million) was \$11,363 million.
- ✗ **Model answered:** For fiscal year ended January 28, 2024, the sum was **\$18,461 million**.
- **Doctor diagnosis:** **Dr. Default** model error
- **Doctor root cause:** model treated the stock-based compensation expense shown in parentheses as a positive amount and added 3,549 to Taiwan revenue instead of applying it as a negative adjustment
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Fabricated number: '\$18,461' not in evidence

Failure 24 — `SEC\_10K\_multi\_neg\_1435\_05` · HIGH confidence

WRONG UNANSWERABLE LABEL

- **Category:** Negative (hallucination traps)
- **Question:** What is the specific dollar amount of noninterest-bearing liabilities that were funding interest-earning assets in 2023?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** The amount was \$731,939 million in 2023.
- **Doctor diagnosis:** **Dr. No** bad test
- **Doctor 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

Complete Failure Listing — Without VERITROOPER (Baseline)

124 confirmed model errors + 13 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 Doctor's diagnosis, and (where available) the Verifier's independent audit verdict.

Failure shape at a glance — Without VERITROOPER (Baseline)

Category	Failures	Model error	Bad test	Review recommended
Negative (hallucination traps)	70	67	3	0
Cross-Reference	43	36	7	0
Precision	15	13	2	0
Exception	4	3	1	0
Cause & Effect	3	3	0	0
Conditional	2	2	0	0
TOTAL	137	124	13	0

Severity mix (confirmed errors): 109 HIGH, 15 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\_0176\_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:** Exception
- **Question:** What portion of Net Interest Income (NII) is excluded to arrive at the \$92.6 billion figure that was flat year-over-year?
- ✓ **Ground truth:** The NII from the Markets business is excluded.
- ✗ **Model answered:** The portion excluded is **Markets net interest income**.

For 2025:

- Total NII: **\$95.443 billion**
- NII excluding Markets: **\$92.591 billion**

So, the excluded amount was:

\$95.443B – \$92.591B = \$2.852 billion

Approximately **\$2.9 billion of Markets NII** was excluded to arrive at the \$92.6 billion figure.

- **Doctor diagnosis:** **Dr. Default** model error
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Fabricated number: '443B' not in evidence

Failure 2 — `SEC\_10K\_multi\_neg\_0851\_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:** Which year's data is represented in the first set of columns for the Transportation sector?

- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** The first set of columns represents fiscal year 2026 - the year ended Jan. 25, 2026.
- **Doctor diagnosis:** **Dr. No** model error
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 3 — `SEC\_10K\_multi\_neg\_0005\_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 was the percentage change in iPad net sales between 2024 and 2023?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** iPad net sales decreased by **6%** from 2023 to 2024.
- **Doctor diagnosis:** **Dr. No** model error
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 4 — `SEC\_10K\_multi\_0344\_02` · HIGH confidence

BAD QUESTION

- **Category:** Cross-Reference
- **Question:** Considering the total client assets were \$7.1 trillion at the end of 2025, what was the approximate value, in billions of dollars, of the total client assets at the end of 2024?
- ✓ **Ground truth:** The total client assets at the end of 2024 were approximately \$5,917 billion. The 2025 value of \$7.1 trillion (\$7,100 billion) represented a 20% increase from 2024, so the 2024 value is calculated as \$7,100 billion divided by 1.20.
- ✗ **Model answered:** The total client assets at the end of 2024 were approximately **\$5,932 billion** (about **\$5.9 trillion**).
- **Doctor diagnosis:** **Dr. Default** bad test
- **Verifier verdict:** bad test (confidence: high)
- **Validator flags:** Fabricated number: '\$5.9' not in evidence

Failure 5 — `t2\_SEC\_10K\_multi\_0765\_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:** By how much does the combined value of Corporate debt securities and Loans exceed the total value of all mortgage-backed securities?
- ✓ **Ground truth:** The combined value of Corporate debt securities (\$463) and Loans (\$759) is \$1,222, which exceeds the total value of mortgage-backed securities (\$771) by \$451.
- ✗ **Model answered:** Using the Level 3 values:
 - Corporate debt securities: \$454 million
 - Loans: \$1,143 million
 - Combined: \$1,597 million
 - Total mortgage-backed securities: \$312 million

Excess = \$1,597 million – \$312 million = **\$1,285 million**.

- **Doctor diagnosis:** **Dr. Default** model error
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Fabricated number: '\$1,597' not in evidence

Failure 6 — `SEC\_10K\_multi\_neg\_0526\_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 was the total dollar amount of credit card net charge-offs for the year ended December 31, 2025?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** The total credit card net charge-offs for the year ended December 31, 2025 were **\$7.67 2 billion**.
- **Doctor diagnosis:** **Dr. No** model error
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 7 — `SEC\_10K\_multi\_0470\_02` · HIGH confidence

EXCEPTION MISSED

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Exception
- **Question:** According to the provided text, the summary of funding sources as of year-end 2025 and 2024 is comprehensive with what stated exception?
- ✓ **Ground truth:** The summary of funding sources is provided with the exception of deposits, which are excluded.
- ✗ **Model answered:** The stated exception is **off-balance sheet obligations**, for which the text says to **refer to Note 28** for additional information.
- **Doctor diagnosis:** **Dr. Default** model error
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 8 — `SEC\_10K\_multi\_neg\_0211\_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 was the operating margin for the period in which the provision for credit losses was \$9,320?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** The provision for credit losses of **\$9,320 million** was in **2023**.

Using **operating margin = pre-provision profit / total net revenue**:

$$\frac{932}{158104} = 44.9\%$$

Operating margin: 44.9%

- **Doctor diagnosis:** **Dr. No** model error
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 9 — `SEC\_10K\_multi\_1387\_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 was the company's Net Income in 2025, in millions of dollars?
- ✓ **Ground truth:** 57,048
- ✗ **Model answered:** The company's Net Income in 2025 was **\$112,010 million**.
- **Doctor diagnosis:** **Dr. Default** model error
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Fabricated number: '\$112,010' not in evidence

Failure 10 — `SEC\_10K\_multi\_0206\_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:** By how much did the net addition to the allowance for loan losses in 2025 fall short of the total year-over-year increase in the allowance for loan losses shown on the balance sheet?
- ✓ **Ground truth:** The net addition to the allowance for loan losses was \$1.4 billion, but the total increase on the balance sheet was \$1,420 million (\$25,765 million minus \$24,345 million). Therefore, the net addition fell short of the total increase by \$20 million.
- ✗ **Model answered:** It did **not** fall short.

Using the stated net addition of **\$4.4 billion** and the year-over-year increase in total allowance for loan losses of **\$1.046 billion** (\$17.557B – \$16.511B), the net addition **exceeded** the balance-sheet increase by **\$3.354 billion**.

- **Doctor diagnosis:** **Dr. Default** model error
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Fabricated number: '557B' not in evidence

Failure 11 — `SEC\_10K\_multi\_neg\_0924\_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 specific balance sheet date mentioned for the basis adjustments associated with de-designated hedged items?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** The balance sheet dates mentioned are **December 31, 2025 and December 31, 2024**.
- **Doctor diagnosis:** **Dr. No** model error
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 12 — `t2\_SEC\_10K\_multi\_0691\_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 value for U.S. Treasury and government agencies securities?
- ✓ **Ground truth:** 234,779
- ✗ **Model answered:** As of December 31, 2025, U.S. Treasury and government agencies securities had a fair value of **\$315.822 billion** in available-for-sale securities.

Including held-to-maturity U.S. Treasury and government agencies securities of **\$126.447 billion**, the combined fair value was **\$442.269 billion**.

- **Doctor diagnosis:** **Dr. Default** model error
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Fabricated number: '\$315.822' not in evidence

Failure 13 — `SEC\_10K\_multi\_neg\_1109\_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 amortized cost basis for the 'Other' category of loan modifications?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** The amortized cost basis for the "Other" category of loan modifications is **\$45 million**.
- **Doctor diagnosis:** **Dr. No** model error
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 14 — `SEC\_10K\_multi\_neg\_0661\_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:** Which regulatory body issued the 'Hedge Accounting Improvements' guidance in November 2025?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** The **Financial Accounting Standards Board (FASB)** issued the "Hedge Accounting Improvements" guidance in November 2025.
- **Doctor diagnosis:** **Dr. No** model error
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 15 — `SEC\_10K\_multi\_neg\_0020\_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:** Which specific domestic and international tax laws were subject to significant judgment by Apple's management in their evaluation of uncertain tax positions?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** Apple's management identified significant judgment in applying complex domestic and international tax laws, specifically including:
 - The TCJA (Tax Cuts and Jobs Act)
 - The allocation of international taxation rights between countries
- **Doctor diagnosis:** **Dr. No** model error
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 16 — `SEC\_10K\_multi\_neg\_1367\_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 target Return on Equity (ROE) for the Commercial & Investment Bank segment?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** The Commercial & Investment Bank segment's ROE is **18%**.
- **Doctor diagnosis:** **Dr. No** model error
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 17 — `SEC\_10K\_multi\_neg\_0010\_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 interest rate on the \$8.0 billion of commercial paper outstanding as of September 27, 2025?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** The weighted-average interest rate on the \$8.0 billion of commercial paper outstanding as of September 27, 2025 was **4.19%**.
- **Doctor diagnosis:** **Dr. No** model error
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 18 — `SEC\_10K\_multi\_neg\_0022\_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 maximum number of shares authorized for issuance under the Apple Inc. 2022 Employee Stock Plan (Exhibit 10.6)?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** Approximately **1.3 billion shares** were authorized for issuance under the Apple Inc. 2022 Employee Stock Plan.

- **Doctor diagnosis:** **Dr. No** model error
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 19 — `t2\_SEC\_10K\_multi\_0882\_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 combined total of OTC credit contracts and OTC-cleared foreign exchange contracts?
- ✓ **Ground truth:** The combined total is 15,745, which is the sum of 14,848 for OTC credit contracts and 897 for OTC-cleared foreign exchange contracts.
- ✗ **Model answered:** Assuming you mean the 2025 net derivative receivables:
 - OTC credit contracts: \$290 million
 - OTC-cleared foreign exchange contracts: \$55 million

Combined total: \$345 million.

- **Doctor diagnosis:** **Dr. Default** model error
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Fabricated number: '\$345' not in evidence

Failure 20 — `SEC\_10K\_multi\_neg\_0452\_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:** On what specific date in 2025 did the Firm issue its new series of non-cumulative preferred stock?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** February 4, 2025.
- **Doctor diagnosis:** **Dr. No** model error
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 21 — `t2\_SEC\_10K\_multi\_1226\_01` · HIGH confidence

CONDITIONAL LOGIC

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Conditional
- **Question:** Under what circumstances would JPMorganChase assess an asset for impairment?
- ✓ **Ground truth:** Impairment is assessed when events or changes in circumstances indicate that the carrying value of an asset may not be fully recoverable.
- ✗ **Model answered:** JPMorganChase would assess impairment when conditions could cause a reporting unit's estimated fair value to decline, such as:
 - Unanticipated declines in business performance
 - Increases in credit losses
 - Increases in capital requirements

- Deterioration in economic or market conditions
- Adverse regulatory or legislative changes
- Increases in the estimated market cost of equity

These circumstances could result in a material impairment charge related to goodwill.

- **Doctor diagnosis:** **Dr. Default** model error
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 22 — `SEC\_10K\_multi\_0666\_02` · 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 was the fair value of federal funds purchased and securities loaned or sold under repurchase agreements in the more recent of the two periods shown?
- ✓ **Ground truth:** 360,194
- ✗ **Model answered:** In the more recent period shown, December 31, 2025, federal funds purchased and securities loaned or sold under repurchase agreements were \$442.396 billion.
- **Doctor diagnosis:** **Dr. Default** model error
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Fabricated number: '\$442.396' not in evidence

Failure 23 — `SEC\_10K\_multi\_neg\_0010\_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 was the total amount of the provision for income taxes related to the State Aid Decision that caused the year-over-year decrease in 2025?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** The year-over-year decrease in the provision for income taxes related to the State Aid Decision in 2025 was **\$10.7 billion**.
- **Doctor diagnosis:** **Dr. No** model error
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 24 — `SEC\_10K\_multi\_neg\_0010\_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 title of the company's chief operating decision maker (CODM) as required by ASU 2023-07?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** The company's chief operating decision maker (CODM) is the **Chairman of the Board of Directors and Chief Executive Officer**.
- **Doctor diagnosis:** **Dr. No** model error
- **Verifier verdict:** model error (confidence: high)

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

Failure 25 — `SEC\_10K\_multi\_neg\_1061\_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:** According to Note 34, what was the total acquisition cost for First Republic?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** According to Note 34, the total acquisition cost for First Republic was **\$67.865 billion**.
- **Doctor diagnosis:** **Dr. No** model error
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer; Fabricated number: '\$67.865' not in evidence

Failure 26 — `SEC\_10K\_multi\_neg\_0968\_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 was the total revenue from deposit-related fees for the year ended December 31, 2025?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** The total revenue from deposit-related fees for the year ended December 31, 2025 was **\$6.876 billion**.
- **Doctor diagnosis:** **Dr. No** model error
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer; Fabricated number: '\$6.876' not in evidence

Failure 27 — `SEC\_10K\_multi\_neg\_0001\_08` · 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 specific form number for submitting an Interactive Data File pursuant to Rule 405 of Regulation S-T?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** The evidence identifies the Interactive Data File as **Exhibit 101** (e.g., **101.INS**, **101.SCH**, etc.) pursuant to Rule 405 of Regulation S-T.
- **Doctor diagnosis:** **Dr. No** model error
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 28 — `SEC\_10K\_multi\_0340\_01` · HIGH confidence

CROSS-REFERENCE MISS

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Cross-Reference

- **Question:** What is the dollar value difference between the 'Allowance for loan losses' in the second period and the 'Total allowance for credit losses' in the third period?
- ✓ **Ground truth:** The difference is \$122. This is calculated by subtracting the 'Allowance for loan losses' of \$539 in the second period from the 'Total allowance for credit losses' of \$661 in the third period.
- ✗ **Model answered:** The difference is **\$1,881 million**, or **\$1.881 billion**.

Calculation: **\$9,175 million – \$7,294 million = \$1,881 million.**

- **Doctor diagnosis:** **Dr. Default** model error
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 29 — `SEC\_10K\_multi\_0334\_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 was the total change in the provision for credit losses from 2024 to 2025, and what was the primary driver for the 2025 provision?
- ✓ **Ground truth:** The provision for credit losses was \$97 million in 2025, compared to a net benefit of \$68 million in the prior year, resulting in a total change of \$165 million. The 2025 provision was largely driven by the impact of a charge-off related to a client-specific exposure.
- ✗ **Model answered:** The provision for credit losses increased by \$3.534 billion, from \$10.678 billion in 2024 to \$14.212 billion in 2025.

The primary driver of the 2025 provision was the consumer provision, especially Card Services, reflecting loan growth, higher net charge-offs, and a \$2.2 billion allowance addition related to the Apple Card transaction.

- **Doctor diagnosis:** **Dr. Default** model error
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Fabricated number: '\$3.534' not in evidence

Failure 30 — `SEC\_10K\_multi\_neg\_1095\_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 percentage of the noninvestment-grade revolving loans were charged off in 2025?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** For 2025 **Commercial and industrial** loans, noninvestment-grade revolving loans were:
 - \$46,721 million within the revolving period
 - \$100 million converted to term loans
 - Total: \$46,821 million

Gross charge-offs on those revolving loans were:

- \$461 million + \$8 million = \$469 million

Percentage charged off:

$\frac{469}{46,821} \approx 1.0\%$

\]

Answer: Approximately 1.0%.

- **Doctor diagnosis:** **Dr. No** model error
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 31 — `SEC\_10K\_multi\_neg\_1111\_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 specific interest rate reduction, in basis points, for loans categorized under 'Other' modifications?
- **Ground truth:** UNANSWERABLE
- **Model answered:** For the 2025 retained credit card loans categorized as "Other" modifications, the weighted-average contractual interest rate was reduced from 22.75% to 8.09%, a reduction of 14.66 percentage points, or 1,466 basis points.
- **Doctor diagnosis:** **Dr. No** model error
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer; Fabricated number: '1,466' not in evidence

Failure 32 — `t2\_SEC\_10K\_multi\_0272\_01` · HIGH confidence

BAD QUESTION

- **Category:** Cross-Reference
- **Question:** If the total taxable-equivalent adjustments for 2025 were reclassified as Investment Banking revenue for that year, what would the new Investment Banking revenue be as a percentage of the original Total Banking & Payments revenue for 2025?
- **Ground truth:** The new Investment Banking revenue would be approximately 35.2% of the original Total Banking & Payments revenue. The taxable-equivalent adjustments for 2025 were \$2.9 billion. Adding this to the 2025 Investment Banking revenue of \$10,198 million (\$10.198 billion) results in a new total of \$13.098 billion. The original Total Banking & Payments revenue for 2025 was \$37,136 million (\$37.136 billion). Therefore, the new Investment Banking revenue would be \$13.098 billion / \$37.136 billion, which is approximately 35.2%.
- **Model answered:** New Investment Banking revenue = \$10,198 million + \$2,900 million = \$13,098 million.

As a percentage of original 2025 Total Banking & Payments revenue:

$$\$13,098 / \$37,136 \times 100 = 35.3\%$$

Answer: 35.3%

- **Doctor diagnosis:** **Dr. Default** bad test
- **Verifier verdict:** bad test (confidence: high)
- **Validator flags:** Fabricated number: '\$2,900' not in evidence

Failure 33 — `t2\_SEC\_10K\_multi\_1096\_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, in millions, of gross charge-offs for term loans originated from 2020 through 2023?
- ✓ **Ground truth:** The total gross charge-offs for term loans originated from 2020 through 2023 is \$117 million, calculated by summing the charge-offs for 2023 (\$38M), 2022 (\$3M), 2021 (\$36M), and 2020 (\$40M).
- ✗ **Model answered:** The question is ambiguous because multiple loan categories are shown. For term loans originated from 2020 through 2023, gross charge-offs were:
 - Secured by real estate: \$172 million
 - Commercial and industrial: \$175 million
 - Other: \$117 million
 - Residential real estate: \$2 million
- **Doctor diagnosis:** **Dr. Default** model error
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Fabricated number: '\$175' not in evidence

Failure 34 — `SEC\_10K\_multi\_neg\_1171\_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 was the total dollar amount of amortization reported in income tax expense for the alternative energy programs for the year ended December 31, 2025?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** \$224 million.
- **Doctor diagnosis:** **Dr. No** model error
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 35 — `SEC\_10K\_multi\_neg\_0237\_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 percentage of the \$4,669 million in asset management fees for 2025 came from J.P. Morgan Wealth Management?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** 100%.

Calculation: \$4,669 million ÷ \$4,669 million = 100%.

- **Doctor diagnosis:** **Dr. No** model error
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 36 — `SEC\_10K\_multi\_neg\_0866\_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 notional amount for interest rate derivative contracts?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** The total notional amount for interest rate derivative contracts was **\$29,536 billion** as of **December 31, 2025**.

For comparison, it was **\$29,660 billion** as of **December 31, 2024**.

- **Doctor diagnosis:** **Dr. No** model error
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 37 — `SEC\_10K\_multi\_neg\_1109\_01` · HIGH confidence

BAD QUESTION

- **Category:** Negative (hallucination traps)
- **Question:** For loans receiving multiple modifications including principal forgiveness, what was the specific deadline for recapturing the deferred payment amounts?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** For loans with multiple modifications including principal forgiveness, the deferred payment amounts were recaptured **at maturity**.
- **Doctor diagnosis:** **Dr. No** bad test
- **Verifier verdict:** bad test (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 38 — `SEC\_10K\_multi\_neg\_0211\_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 currency are the dollar amounts in the table presented in (e.g., USD, CAD, AUD)?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** USD - U.S. dollars, in millions.
- **Doctor diagnosis:** **Dr. No** model error
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 39 — `SEC\_10K\_multi\_0272\_02` · HIGH confidence

PRECISION ERROR

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Precision
- **Question:** What was the value of taxable-equivalent adjustments for the year ended December 31, 2024?
- ✓ **Ground truth:** \$2.8 billion
- ✗ **Model answered:** For the year ended December 31, 2024, the fully taxable-equivalent adjustments were:
 - Net interest income: \$477 million
 - Noninterest revenue / other income: \$2,560 million
 - Total net revenue adjustment: \$3,037 million
- **Doctor diagnosis:** **Dr. Default** model error
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 40 — `SEC\_10K\_multi\_0687\_02` · 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 value of Equity securities, in millions of U.S. dollars?
- ✓ **Ground truth:** 131,728
- ✗ **Model answered:** As of December 31, 2025, total equity securities were **\$2,633 million** combined:
 - **\$2,589 million** measured at fair value
 - **\$44 million** under the measurement alternative
- **Doctor diagnosis:** **Dr. Default** model error
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Fabricated number: '\$2,633' not in evidence

Failure 41 — `SEC\_10K\_multi\_0280\_01` · HIGH confidence

PRECISION ERROR

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Precision
- **Question:** What was the reported revenue for Advisory fees in 2025?
- ✓ **Ground truth:** \$3.5 billion
- ✗ **Model answered:** The reported revenue for Advisory fees in 2025 was **\$3,503 million**.
- **Doctor diagnosis:** **Dr. Default** model error
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 42 — `SEC\_10K\_multi\_neg\_0001\_01` · HIGH confidence

BAD QUESTION

- **Category:** Negative (hallucination traps)
- **Question:** What is the Commission File Number for Apple's Transition Report, should one be filed?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** Apple's Commission File Number is **001-36743**.
- **Doctor diagnosis:** **Dr. No** bad test
- **Verifier verdict:** bad test (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 43 — `SEC\_10K\_multi\_0873\_01` · HIGH confidence

BAD QUESTION

- **Category:** Precision
- **Question:** What is the total fair value of trading assets and liabilities for Interest rate, including both Level 1 and Level 2 assets?
- ✓ **Ground truth:** \$274,228
- ✗ **Model answered:** For Interest rate:
 - Trading assets (derivative receivables), Level 1 + Level 2: \$4,934 million + \$282,019 million = \$286,953 million

- Trading liabilities (derivative payables), Level 1 + Level 2: \$2,071 million + \$253,078 million = \$255,149 million

Combined total: \$542,102 million.

- **Doctor diagnosis:** **Dr. Default** bad test
- **Verifier verdict:** bad test (confidence: high)
- **Validator flags:** Fabricated number: '\$542,102' not in evidence

Failure 44 — `SEC\_10K\_multi\_neg\_0047\_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 expiration date for the research and other tax credit carryforwards listed as a deferred tax asset?
- **Ground truth:** UNANSWERABLE
- **Model answered:** The research tax credit carryforwards expire as follows:
 - Federal research tax credits: begin to expire in fiscal year 2027.
 - California state research tax credits: may be carried forward indefinitely.
 - Other state research tax credits: begin to expire in fiscal year 2028.
- **Doctor diagnosis:** **Dr. No** model error
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 45 — `SEC\_10K\_multi\_neg\_0526\_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 fee for underwriting services in the wholesale credit portfolio?
- **Ground truth:** UNANSWERABLE
- **Model answered:** Underwriting fees reported were:
 - **Debt underwriting fees:** \$4.5 billion
 - **Equity underwriting fees:** \$1.7 billion

Combined underwriting fees were \$6.2 billion.

- **Doctor diagnosis:** **Dr. No** model error
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer; Fabricated number: '\$6.2' not in evidence

Failure 46 — `SEC\_10K\_multi\_neg\_0924\_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 was the total change in Other Comprehensive Income (OCI) for interest rate contracts for the year ended December 31, 2024?

- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** The total change in OCI for interest rate contracts for the year ended December 31, 2024 was a loss of **\$935 million**.
- **Doctor diagnosis:** **Dr. No** model error
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 47 — `SEC\_10K\_multi\_0334\_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:** By what percentage did the combined revenue from Asset Management and Global Private Bank exceed the total Noninterest expense in 2025?
- ✓ **Ground truth:** The combined revenue from Asset Management (\$11.7 billion) and Global Private Bank (\$12.4 billion) was \$24.1 billion. Total Noninterest expense was \$15.3 billion. The combined revenue exceeded the expense by \$8.8 billion, which is approximately 57.5% of the expense.
- ✗ **Model answered:** In 2025, combined revenue from Asset Management and Global Private Bank was:

\$11.700B + \$12.373B = \$24.073B

Total noninterest expense was \$15.332B.

Percentage by which revenue exceeded expense:

$$\frac{24.073 - 15.332}{15.332} \times 100 \approx 57.0\%$$

Answer: The combined revenue exceeded total noninterest expense by approximately 57%.

- **Doctor diagnosis:** **Dr. Default** model error
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Fabricated number: '700B' not in evidence

Failure 48 — `t2\_SEC\_10K\_multi\_1367\_01` · HIGH confidence

CAUSE-EFFECT ERROR

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Cause & Effect
- **Question:** What was the effect of lower average interest rates during the year ended December 31, 2025?
- ✓ **Ground truth:** As a result of lower average interest rates, the cost of funding for assets and the funding benefit earned for liabilities generally decreased, which resulted in a lower cost of funds for loans and Markets activities.
- ✗ **Model answered:** The evidence does not state an effect from lower average interest rates for the year ended December 31, 2025. It states that interest income decreased by \$202 million, primarily due to lower average investment balances, and interest expense decreased by \$2 million, primarily due to lower average short-term debt balances.
- **Doctor diagnosis:** **Dr. Default** model error
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 49 — `SEC\_10K\_multi\_0245\_02` · 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 was the year-over-year increase in the net addition to the allowance for credit losses from 2024 to 2025, and how much of the 2025 addition was due to the Apple Card transaction?
- ✓ **Ground truth:** The net addition to the allowance for credit losses increased by \$1.2 billion, from \$2.0 billion in the prior year (2024) to \$3.2 billion in 2025. The Apple Card transaction accounted for \$2.2 billion of the 2025 addition.
- ✗ **Model answered:** The net addition to the allowance for credit losses increased by **\$2.4 billion**, from **\$2.0 billion in 2024 to \$4.4 billion in 2025**.

Of the 2025 addition, **\$2.2 billion** was related to the **Apple Card transaction**.

- **Doctor diagnosis:** **Dr. Default** model error
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Fabricated number: '\$2.4' not in evidence

Failure 50 — `SEC\_10K\_multi\_0687\_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 value of Corporate debt securities, in millions of U.S. dollars?
- ✓ **Ground truth:** 33,981
- ✗ **Model answered:** The total value of Corporate debt securities is **\$17,418 million**.
- **Doctor diagnosis:** **Dr. Default** model error
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Fabricated number: '\$17,418' not in evidence

Failure 51 — `SEC\_10K\_multi\_neg\_1432\_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 average balance of noninterest-earning assets?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** For 2025, the average balance of total noninterest-earning assets was **\$597.265 billion**.

The line item "**All other noninterest-earning assets**" was **\$233.375 billion**.

- **Doctor diagnosis:** **Dr. No** model error
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer; Fabricated number: '\$597.265' not in evidence

Failure 52 — `SEC\_10K\_multi\_neg\_1109\_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 was the total dollar amount of interest income lost due to the interest rate reductions on loans with multiple modifications?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** Approximately **\$13.6 million**.

Calculation:

- \$86 million × 10.60% = \$9.1 million
- \$67 million × 6.72% = \$4.5 million
- Total = **\$13.6 million** in reduced interest income.
- **Doctor diagnosis:** **Dr. No** model error
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer; Fabricated number: '10.60%' not in evidence

Failure 53 — `SEC\_10K\_multi\_neg\_0237\_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 were the total noninterest expenses for the Consumer & Community Banking division in 2023?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** The Consumer & Community Banking division's total noninterest expense in 2023 was **\$34.819 billion**.
- **Doctor diagnosis:** **Dr. No** model error
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 54 — `t2\_SEC\_10K\_multi\_1096\_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, in millions, of noninvestment-grade term loans originated in 2022 and 2023 combined?
- ✓ **Ground truth:** The total value is \$13,166 million, which is the sum of the \$5,918 million from 2022 and the \$7,248 million from 2023.
- ✗ **Model answered:** As of December 31, 2025, across the listed wholesale loan categories, noninvestment-grade term loans originated in 2022 and 2023 combined totaled **\$36,867 million**.

Calculation:

- Secured by real estate: \$12,444 + \$4,392 = \$16,836 million
- Commercial and industrial: \$5,600 + \$5,927 = \$11,527 million
- Other: \$4,079 + \$4,425 = \$8,504 million

- Total: \$16,836 + \$11,527 + \$8,504 = **\$36,867 million**
- **Doctor diagnosis:** **Dr. Default** model error
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Fabricated number: '\$36,867' not in evidence

Failure 55 — `SEC\_10K\_multi\_1053\_00` · HIGH confidence

BAD QUESTION

- **Category:** Cross-Reference
- **Question:** Considering the loan presentation method described in the first paragraph and the data in the table for the year ended December 31, 2025, what was the total amount, in millions, of loans purchased and sold, excluding those for which the fair value option has been elected?
- ✓ **Ground truth:** The total amount of loans purchased was \$2,452 million and the total amount of loans sold was \$55,583 million. These loans are presented net of unamortized discounts and premiums and net deferred loan fees or costs.
- ✗ **Model answered:** For the year ended December 31, 2025, excluding loans for which the fair value option was elected:
 - Loans purchased: \$2,452 million
 - Loans sold: \$55,583 million

Total purchased and sold: \$58,035 million.

- **Doctor diagnosis:** **Dr. Default** bad test
- **Verifier verdict:** bad test (confidence: high)
- **Validator flags:** Fabricated number: '\$58,035' not in evidence

Failure 56 — `SEC\_10K\_multi\_neg\_0913\_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 were the total dollar amounts of the cumulative fair value hedge basis adjustments recorded on the Consolidated balance sheets as of December 31, 2025 and 2024?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** The total cumulative fair value hedge basis adjustments recorded on the Consolidated balance sheets were:
 - **December 31, 2025: \$(6.101) billion**
 - **December 31, 2024: \$(16.712) billion**

These totals are the sum of the "Total" basis adjustments for AFS investment securities, long-term debt, and beneficial interests issued by consolidated VIEs.

- **Doctor diagnosis:** **Dr. No** model error
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer; Fabricated number: '6.101' not in evidence

Failure 57 — `SEC\_10K\_multi\_neg\_1435\_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 was the total net change in interest income for the period of 2025 versus 2024 for all interest-earning assets combined?
- **Ground truth:** UNANSWERABLE
- **Model answered:** For 2025 versus 2024, total interest income on all interest-earning assets decreased by **\$644 million** on a taxable-equivalent basis, from **\$194.410 billion** in 2024 to **\$193.766 billion** in 2025.
- **Doctor diagnosis:** **Dr. No** model error
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer; Fabricated number: '\$194.410' not in evidence

Failure 58 — `t2\_SEC\_10K\_multi\_1405\_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 intangible assets acquired from First Republic that are amortized over approximately 7 years?
- **Ground truth:** The total value is \$1.48 billion, which is the sum of the \$1.3 billion core deposit intangible and the \$180 million customer relationship intangibles.
- **Model answered:** The total value was **\$1.455 billion** in core deposit and customer relationship intangibles, amortized over approximately **7 years**.
- **Doctor diagnosis:** **Dr. Default** model error
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Fabricated number: '\$1.455' not in evidence

Failure 59 — `t2\_SEC\_10K\_multi\_0253\_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:** Using the net charge-off rate for Card Services and its corresponding net charge-offs, what was the approximate average loan balance for that segment?
- **Ground truth:** The average loan balance for Card Services was approximately \$231,964, calculated by dividing the net charge-offs of \$7,678 by the net charge-off rate of 3.31% (or 0.0331).
- **Model answered:** Using Card Services net charge-offs of **\$7.678 billion** and a net charge-off rate of **3.31%**:

$$\text{Average loan balance} \approx \frac{7.678}{0.0331} = 232 \text{ billion}$$

So, the approximate average loan balance for Card Services was **about \$232 billion**.

- **Doctor diagnosis:** **Dr. Default** model error
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Fabricated number: '\$7.678' not in evidence

Failure 60 — `SEC\_10K\_multi\_neg\_0661\_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 form number used to document the election of the new hedge accounting guidance?
- **Ground truth:** UNANSWERABLE
- **Model answered:** Form 10-K.
- **Doctor diagnosis:** **Dr. No** model error
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 61 — `t2\_SEC\_10K\_multi\_0801\_01` · HIGH confidence

CROSS-REFERENCE MISS

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Cross-Reference
- **Question:** Using the values from the first numerical column for each item, what is the sum of 'Deposits' and 'Federal funds purchased and securities loaned or sold under repurchase agreements'?
- **Ground truth:** The sum is \$2,620.6, which is the total of \$2,538.4 for 'Deposits' and \$82.2 for 'Federal funds purchased and securities loaned or sold under repurchase agreements'.
- **Model answered:** \$2,559,320 million + \$442,396 million = \$3,001,716 million.
- **Doctor diagnosis:** **Dr. Default** model error
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 62 — `SEC\_10K\_multi\_0344\_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:** By what percentage did the combined total of Fixed income and Equity assets under management increase from the end of 2024 to the end of 2025?
- **Ground truth:** The combined total of Fixed income and Equity assets under management increased by approximately 21.0% from the end of 2024 to the end of 2025. The 2024 total was \$1,979 billion (\$851 billion + \$1,128 billion) and the 2025 total was \$2,398 billion (\$998 billion + \$1,400 billion).
- **Model answered:** The combined Fixed Income and Equity amount increased by **about 16.1%** from the end of 2024 to the end of 2025.

Calculation:

- 2025: \$18,322B + \$17,954B = \$36,276B
- 2024: \$16,409B + \$14,848B = \$31,257B
- Increase: \$5,019B
- Percentage increase: $\$5,019B / \$31,257B \times 100 = 16.1\%$
- **Doctor diagnosis:** **Dr. Default** model error
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Fabricated number: '322B' not in evidence

Failure 63 — `SEC\_10K\_multi\_neg\_0866\_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 was the net gain or loss realized from the total derivative positions of \$50,642?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** The net amount realized was a **gain of \$212 million** from derivative instruments.
- **Doctor diagnosis:** **Dr. No** model error
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 64 — `SEC\_10K\_multi\_neg\_0972\_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:** According to Note 28, what is the total value of the firm's lending-related commitments?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** According to Note 28, the firm's total lending-related commitments were **\$1,817.307 billion** (about **\$1.817 trillion**) as of December 31, 2025.

The related **carrying value** was **\$6.278 billion**.

- **Doctor diagnosis:** **Dr. No** model error
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer; Fabricated number: '\$1,817.307' not in evidence

Failure 65 — `SEC\_10K\_multi\_neg\_1109\_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 was the total amortized cost basis for all Commercial and industrial loans at the end of 2025?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** The total amortized cost basis for Commercial and industrial loans at the end of 2025 was **\$2.462 billion**.
- **Doctor diagnosis:** **Dr. No** model error
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 66 — `SEC\_10K\_multi\_0195\_02` · HIGH confidence

EXCEPTION MISSED

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Exception

- **Question:** What is the only category of provision for credit losses listed in the table that is not part of the consumer or wholesale segments?
- ✓ **Ground truth:** Investment securities.
- ✗ **Model answered:** Corporate.
- **Doctor diagnosis:** **Dr. Default** model error
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 67 — `SEC\_10K\_multi\_neg\_0871\_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 notional value of the derivative contracts included in these fair value calculations?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** The notional value of derivatives designated and qualified as fair value hedges was:
 - \$13.674 billion as of December 31, 2025
 - \$12.628 billion as of December 31, 2024
- **Doctor diagnosis:** **Dr. No** model error
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer; Fabricated number: '\$12.628' not in evidence

Failure 68 — `SEC\_10K\_multi\_neg\_1435\_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 specific date in 2025 by which the Form 10-K must be filed with the SEC?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** February 26, 2025.
- **Doctor diagnosis:** **Dr. No** model error
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 69 — `SEC\_10K\_multi\_neg\_1259\_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:** On what date was the interest rate for Series FF originally set?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** The interest rate for Series FF was originally set on **July 31, 2019**.
- **Doctor diagnosis:** **Dr. No** model error
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 70 — `SEC\_10K\_multi\_neg\_1113\_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 was the total amortized cost basis for all Commercial and Industrial loans for the year ended December 31, 2023?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** The total amortized cost basis for all Commercial and Industrial loans for the year ended December 31, 2023 was **\$1,363 million**.
- **Doctor diagnosis:** **Dr. No** model error
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 71 — `SEC\_10K\_multi\_neg\_0851\_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 unit of measurement for the values listed under the Metals & Mining category?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** The values under the Metals & Mining category are measured **in millions of dollars**.
- **Doctor diagnosis:** **Dr. No** model error
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 72 — `SEC\_10K\_multi\_1256\_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 annual interest payment for all fixed-rate series that mature in 2026?
 - ✓ **Ground truth:** The total annual interest payment for the fixed-rate series maturing in 2026 (Series JJ, LL, and MM) is \$1,337.52, calculated by summing their individual annual interest payments of \$455.00, \$462.52, and \$420.00.
 - ✗ **Model answered:** Assuming "mature in 2026" refers to the fixed-rate preferred series with 2026 earliest redemption dates:
 - Series JJ: $\$1.5\text{B} \times 4.550\% = \68.25M
 - Series LL: $\$1.85\text{B} \times 4.625\% = \85.56M
 - Series MM: $\$2.0\text{B} \times 4.200\% = \84.00M
- Total annual payment: approximately **\$237.8 million**.
- **Doctor diagnosis:** **Dr. Default** model error
 - **Verifier verdict:** model error (confidence: high)
 - **Validator flags:** Fabricated number: '5B' not in evidence

Failure 73 — `t2\_SEC\_10K\_multi\_0765\_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. GSEs and government agencies' mortgage-backed securities combined with the value of Non-U.S. government debt securities?
- ✓ **Ground truth:** The total value is \$914, calculated by adding the \$759 for U.S. GSEs and government agencies' mortgage-backed securities to the \$155 for Non-U.S. government debt securities.
- ✗ **Model answered:** Using the December 31, 2025 fair values:
 - U.S. GSEs and government agencies mortgage-backed securities: \$158,141 million
 - Non-U.S. government debt securities: \$124,684 million

Combined total = \$282,825 million, or \$282.825 billion.

- **Doctor diagnosis:** **Dr. Default** model error
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Fabricated number: '\$282.825' not in evidence

Failure 74 — `t2\_SEC\_10K\_multi\_1339\_01` · HIGH confidence

CROSS-REFERENCE MISS

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Cross-Reference
- **Question:** Using the fifth numerical column, what is the total for 'Unsettled resale and securities borrowed agreements' and 'Unsettled repurchase and securities loaned agreements' combined?
- ✓ **Ground truth:** The combined total is \$189,967, which is the sum of \$137,072 for 'Unsettled resale and securities borrowed agreements' and \$52,895 for 'Unsettled repurchase and securities loaned agreements'.
- ✗ **Model answered:** \$189,967 million.
- **Doctor diagnosis:** **Dr. Default** model error
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 75 — `SEC\_10K\_multi\_neg\_0001\_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 was the total dollar amount of shares repurchased by Apple during the three months ended September 27, 2025?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** Apple repurchased approximately **\$20.0 billion** of its common stock during the three months ended September 27, 2025.

Calculation from the provided periods:

- \$33.265 million shares × \$210.43 ≈ \$7.0 billion
- \$28.986 million shares × \$224.25 ≈ \$6.5 billion
- \$27.247 million shares × \$238.56 ≈ \$6.5 billion

Total ≈ **\$20.0 billion**.

- **Doctor diagnosis:** **Dr. No** model error
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 76 — `SEC\_10K\_multi\_1078\_02` · HIGH confidence

CONDITIONAL LOGIC

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Conditional
- **Question:** If a home valuation model is used to estimate current property values, what type of valuation estimates must it incorporate?
- ✓ **Ground truth:** It must incorporate nationally recognized home price index valuation estimates.
- ✗ **Model answered:** It must incorporate **fair value estimates of the underlying collateral/property**, based on relevant valuation inputs such as comparable valuations, transaction details, and other empirical market data.
- **Doctor diagnosis:** **Dr. Default** model error
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 77 — `SEC\_10K\_multi\_0600\_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:** Using the table for the allowance for loan losses, what was the total allowance for consumer loans in 2025, and based on the Investment Portfolio Risk Management section, what was the total carrying value of principal investments for the same year?
- ✓ **Ground truth:** The total allowance for consumer loans in 2025 was \$17,477 million, and the total carrying value of principal investments was \$47.0 billion.
- ✗ **Model answered:** The total allowance for consumer loans in 2025 was **\$17.477 billion**.

This equals:

- Consumer, excluding credit card: **\$1.920 billion**
- Credit card: **\$15.557 billion**
- Total consumer: **\$17.477 billion**

The provided Investment Portfolio Risk Management excerpt does **not disclose the total carrying value of principal investments for 2025**.

- **Doctor diagnosis:** **Dr. Default** model error
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Fabricated number: '\$1.920' not in evidence

Failure 78 — `SEC\_10K\_multi\_0871\_00` · HIGH confidence

BAD QUESTION

- **Category:** Precision
- **Question:** As of December 31, 2024, what was the total fair value of trading assets and liabilities designated as hedges, in millions?
- ✓ **Ground truth:** \$2,609
- ✗ **Model answered:** As of December 31, 2024, the total fair value designated as hedges was:

- Trading assets / derivative receivables: \$2,021 million
- Trading liabilities / derivative payables: \$1,353 million

Combined total: \$3,374 million.

- **Doctor diagnosis:** **Dr. Default** bad test
- **Verifier verdict:** bad test (confidence: high)
- **Validator flags:** Fabricated number: '\$3,374' not in evidence

Failure 79 — `SEC\_10K\_multi\_neg\_0450\_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 specific GSIB surcharge percentage used in the Firm's equity allocation methodology?
- **Ground truth:** UNANSWERABLE
- **Model answered:** The Firm's equity allocation methodology uses the GSIB surcharge under rules currently in effect; based on the evidence, the applicable higher GSIB surcharge is **4.5%**.
- **Doctor diagnosis:** **Dr. No** model error
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 80 — `t2\_SEC\_10K\_multi\_0272\_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:** Considering the taxable-equivalent adjustments for 2024 and the compensation expense as a percentage of total net revenue for the same year, what would be the total compensation expense if the total net revenue was exactly 15 times the value of the taxable-equivalent adjustments?
- **Ground truth:** The compensation expense would be \$10.92 billion. The taxable-equivalent adjustments for 2024 were \$2.8 billion. If total net revenue was 15 times this amount, it would be \$42 billion ($15 \times \2.8 billion). The compensation expense as a percentage of total net revenue for 2024 was 26%, so the total compensation expense would be 26% of \$42 billion, which is \$10.92 billion.
- **Model answered:** Using the 2024 taxable-equivalent adjustment for total net revenue of **\$3,037 million** and the 2024 compensation expense ratio of **26%**:
 - Hypothetical total net revenue = $15 \times \$3,037 \text{ million} = \$45,555 \text{ million}$
 - Compensation expense = $26\% \times \$45,555 \text{ million} = \$11,844.3 \text{ million}$

Total compensation expense would be approximately \$11.84 billion.

- **Doctor diagnosis:** **Dr. Default** model error
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Fabricated number: '\$45,555' not in evidence

Failure 81 — `SEC\_10K\_multi\_neg\_1435\_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 specific interest rate for non-U.S. deposits with banks for the year ended December 31, 2024?
- **Ground truth:** UNANSWERABLE
- **Model answered:** The interest rate for non-U.S. deposits with banks for the year ended December 31, 2024 was 3.48%.
- **Doctor diagnosis:** **Dr. No** model error
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 82 — `SEC\_10K\_multi\_0013\_02` · 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 sum of the Fair Value of Mutual funds and the Non-Current Marketable Securities portion of U.S. Treasury securities?
 - **Ground truth:** The sum is \$11,563, which is the Fair Value of Mutual funds (\$617) plus the Non-Current Marketable Securities value of U.S. Treasury securities (\$10,946).
 - **Model answered:** Using the 2025 Apple table:
 - Fair Value of Mutual funds: \$854 million
 - Non-Current Marketable Securities portion of U.S. Treasury securities: \$10,946 million
- Sum = \$11,800 million, or \$11.8 billion.
- **Doctor diagnosis:** **Dr. Default** model error
 - **Verifier verdict:** model error (confidence: high)
 - **Validator flags:** Fabricated number: '\$11.8' not in evidence

Failure 83 — `t2\_SEC\_10K\_multi\_1491\_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:** Using the total notional value of foreign currency derivatives for 2025 and the total notional value of derivatives designated for the foreign currency cash flow hedging program for the same year, what is the total notional value of all other foreign currency derivatives?
- **Ground truth:** The total notional value of all other foreign currency derivatives for 2025 is \$11,368 million.
- **Model answered:** Assuming "total foreign currency derivatives" refers to total foreign exchange contracts:

\$15,595 billion – \$9.760 billion = \$15,585.24 billion

So, the total notional value of all other foreign currency derivatives is \$15,585.24 billion.
- **Doctor diagnosis:** **Dr. Default** model error
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Fabricated number: '\$9.760' not in evidence

Failure 84 — `SEC\_10K\_multi\_1073\_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 value of retained loans for the current period that are NOT from California, New York, or Florida?
- **Ground truth:** The total retained loans for the current period are \$303,531. Subtracting the amounts for California (\$117,500), New York (\$46,378), and Florida (\$21,864) leaves a remaining value of \$117,789.
- **Model answered:** The evidence includes multiple retained-loan portfolios, so the answer depends on which portfolio is meant:
- **Residential real estate retained loans:** \$117.789 billion
= \$303.531B - (\$117.500B + \$46.378B + \$21.864B)
- **Credit card retained loans:** \$170.985 billion
= \$247.797B - (\$38.702B + \$19.488B + \$18.622B)
- **Auto and other retained loans:** \$45.131 billion
= \$65.210B - (\$9.926B + \$4.771B + \$5.382B)
- **Doctor diagnosis:** **Dr. Default** model error
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Fabricated number: '\$117.789' not in evidence

Failure 85 — `SEC\_10K\_multi\_0330\_02` · 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:** Using the total client assets for Asset & Wealth Management and the asset management fees reported for 2025, what was the approximate revenue from asset management fees for every one trillion dollars of client assets?
- **Ground truth:** Asset & Wealth Management has \$7.1 trillion in client assets. In 2025, it generated \$15,494 million in asset management fees. Therefore, the revenue per trillion dollars of client assets was approximately \$2,182 million, or \$2.18 billion.
- **Model answered:** Asset & Wealth Management reported:
- Total client assets: \$7.118 trillion
- 2025 asset management fees: \$15.494 billion

Calculation:

$\$15.494 \text{ billion} \div 7.118 = \text{approximately } \2.18 billion

So, AWM generated approximately **\$2.2 billion in asset management fees for every \$1 trillion of client assets** in 2025.

- **Doctor diagnosis:** **Dr. Default** model error
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Fabricated number: '\$2.2' not in evidence

Failure 86 — `SEC\_10K\_multi\_neg\_0022\_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:** On what page of this Form 10-K can the full text of the Insider Trading Policy (Exhibit 19.1) be found?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** Page 56.
- **Doctor diagnosis:** **Dr. No** model error
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 87 — `SEC\_10K\_multi\_neg\_1435\_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:** According to the 'Changes in net interest income' table, what was the net change attributed to volume for interest-bearing liabilities in 2024 versus 2023?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** The net change attributed to volume for interest-bearing liabilities in 2024 versus 2023 was an increase of approximately **\$11.4 billion**.
- **Doctor diagnosis:** **Dr. No** model error
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer; Fabricated number: '\$11.4' not in evidence

Failure 88 — `t2\_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:** As of December 31, 2025, what was the contractual amount, net of risk participations, for other unfunded commitments to extend credit?
- ✓ **Ground truth:** \$181 million
- ✗ **Model answered:** As of December 31, 2025, the contractual amount, net of risk participations, for other unfunded commitments to extend credit was **\$561.506 billion**.
- **Doctor diagnosis:** **Dr. Default** model error
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Fabricated number: '\$561.506' not in evidence

Failure 89 — `SEC\_10K\_multi\_neg\_1259\_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 payment frequency for the interest on the Series HH notes?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** For Series HH, payments were **semiannual while at the fixed rate**, and would become **quarterly after converting to a floating rate**. The Series HH preferred stock was redeemed on **February 1, 2025**.
- **Doctor diagnosis:** **Dr. No** model error
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 90 — `SEC\_10K\_multi\_0388\_00` · HIGH confidence

CAUSE-EFFECT ERROR

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Cause & Effect
- **Question:** What is identified as the cause for the existence of risk within JPMorgan Chase's business activities?
- ✓ **Ground truth:** Factors such as the economic environment, regulatory or government policy, competitor or market evolution, business decisions, process or judgment error, deliberate wrongdoing, dysfunctional markets, and natural disasters are identified as the drivers that cause risk to exist.
- ✗ **Model answered:** Risk is identified as inherent in JPMorgan Chase's business activities-i.e., the Firm's lines of business, Corporate activities, functions, and regions can give rise to risk.
- **Doctor diagnosis:** **Dr. Default** model error
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 91 — `SEC\_10K\_multi\_neg\_0851\_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 title of the column containing the value 3,051 for Securities Firms?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** The column title is **Derivative receivables**.
- **Doctor diagnosis:** **Dr. No** model error
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 92 — `SEC\_10K\_multi\_neg\_1259\_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 total outstanding principal for all series combined?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** The total outstanding amount for all preferred stock series combined was approximately **\$20.1 billion** at December 31, 2025, based on the aggregate liquidation value.
- **Doctor diagnosis:** **Dr. No** model error
- **Verifier verdict:** model error (confidence: high)

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

Failure 93 — `SEC\_10K\_multi\_1370\_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:** For the year ending December 31, 2025, for the Corporate segment, what is the difference between the Total noninterest expense and the Average equity?
- **Ground truth:** The difference is -\$668 million. This is calculated by taking the Corporate segment's Total noninterest expense of \$15,332 million and subtracting its Average equity of \$16,000 million.
- **Model answered:** For the Corporate segment in 2025:
 - Total noninterest expense: \$1,825 million
 - Average equity: \$111,254 million

Difference: **\$109,429 million.**

- **Doctor diagnosis:** **Dr. Default** model error
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Fabricated number: '\$109,429' not in evidence

Failure 94 — `t2\_SEC\_10K\_multi\_0882\_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:** By how much does the total for OTC foreign exchange contracts exceed the total for all credit contracts combined?
- **Ground truth:** The total for OTC foreign exchange contracts (175,485) exceeds the total for all credit contracts (17,294) by 158,191.
- **Model answered:** The evidence does not specify whether you mean derivative receivables or payables.
 - For 2025 gross derivative receivables: OTC foreign exchange contracts exceed total credit contracts by **\$167,164 million** (\$180,120 million – \$12,956 million).
 - For 2025 gross derivative payables: OTC foreign exchange contracts exceed total credit contracts by **\$158,191 million** (\$175,485 million – \$17,294 million).
- **Doctor diagnosis:** **Dr. Default** model error
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Fabricated number: '\$167,164' not in evidence

Failure 95 — `SEC\_10K\_multi\_neg\_0452\_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:** According to the Federal Reserve's TLAC rule, what is the penalty for failing to maintain the minimum level of eligible long-term debt?
- **Ground truth:** UNANSWERABLE

- **✗ Model answered:** Failure to maintain the required TLAC level results in limitations on capital distributions, including dividends and common share repurchases, and on discretionary bonus payments for certain executive officers.
- **Doctor diagnosis:** **Dr. No** model error
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 96 — `t2\_SEC\_10K\_multi\_0112\_01` · HIGH confidence

CAUSE-EFFECT ERROR

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Cause & Effect
- **Question:** What was the direct consequence of a non-U.S. subsidiary's client needing to renew travel documentation for their children?
- **✓ Ground truth:** The payments made for the renewal of travel documentation were therefore exempt transactions pursuant to 31 C.F.R. 560.219(d).
- **✗ Model answered:** A non-U.S. subsidiary processed three payments, each about USD 130, for the client to the Iranian Embassy in London, and the Firm charged no fee.
- **Doctor diagnosis:** **Dr. Default** model error
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 97 — `SEC\_10K\_multi\_0547\_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:** By how much does the second column value for the Insurance sector exceed the second column value for the Metals & Mining sector?
- **✓ Ground truth:** The second column value for the Insurance sector (17,847) exceeds that of the Metals & Mining sector (7,373) by 10,474.
- **✗ Model answered:** The Insurance sector's second-column value exceeds Metals & Mining's by **\$10,052** million.
- **Doctor diagnosis:** **Dr. Default** model error
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Fabricated number: '\$10,052' not in evidence

Failure 98 — `SEC\_10K\_multi\_0094\_01` · HIGH confidence

BAD QUESTION

- **Category:** Cross-Reference
- **Question:** Besides the market where JPMorganChase's common stock is traded, what two indices are mentioned for a five-year performance comparison ending on December 31, 2025?
- **✓ Ground truth:** JPMorganChase's common stock is traded on the New York Stock Exchange, and the five-year performance comparison ending December 31, 2025, includes the KBW Bank Index and the S&P Financials Index.
- **✗ Model answered:** The evidence mentions three indices, not two: the S&P 500 Index, the KBW Bank Index, and the S&P Financials Index.
- **Doctor diagnosis:** **Dr. Default** bad test
- **Verifier verdict:** bad test (confidence: high)

- **Validator flags:** (none)

Failure 99 — `SEC\_10K\_multi\_1405\_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 value of the core deposit intangible and customer relationship intangibles, and by how much does this combined value exceed the total of tax-oriented investment liabilities and lease liabilities mentioned in paragraph (d)?
- ✓ **Ground truth:** The core deposit intangible is valued at \$1.3 billion and the customer relationship intangibles are valued at \$180 million, for a total of \$1,480 million. The tax-oriented investment liabilities are \$669 million and lease liabilities are \$748 million, totaling \$1,417 million. The combined intangible value exceeds the combined liabilities by \$63 million (\$1,480 million - \$1,417 million).
- ✗ **Model answered:** The total value of the core deposit and customer relationship intangibles was **\$1.455 billion**.

The tax-oriented investment liabilities and lease liabilities totaled:

- \$669 million + \$748 million = **\$1.417 billion**

So, the intangibles exceeded those liabilities by:

- \$1.455 billion - \$1.417 billion = **\$38 million**.
- **Doctor diagnosis:** **Dr. Default** model error
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Fabricated number: '\$1.455' not in evidence

Failure 100 — `SEC\_10K\_multi\_1095\_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 was the total value, in millions, of investment-grade term loans originated in 2023?
- ✓ **Ground truth:** \$7,939
- ✗ **Model answered:** \$20,166 million.
- **Doctor diagnosis:** **Dr. Default** model error
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Fabricated number: '\$20,166' not in evidence

Failure 101 — `SEC\_10K\_multi\_neg\_0871\_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 was the total fair value of trading assets and liabilities as of December 31, 2023?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** As of December 31, 2023, the Level 3 fair value was:
- Trading assets - debt and equity instruments: \$2.373 billion

- Trading liabilities - debt and equity instruments: \$37 million

Total combined fair value: \$2.410 billion.

- **Doctor diagnosis:** **Dr. No** model error
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 102 — `SEC\_10K\_multi\_0801\_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 combined total of 'Deposits' and 'Short-term borrowings' from the first set of five columns?
- ✓ **Ground truth:** The combined total is \$2,571.1, which is the sum of \$2,538.8 for Deposits and \$32.3 for Short-term borrowings.
- ✗ **Model answered:** The combined total is **\$35,555 million**.

Calculation: **Short-term borrowings (\$14,828 million) + Deposits (\$20,727 million) = \$35,555 million.**

- **Doctor diagnosis:** **Dr. Default** model error
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Fabricated number: '\$35,555' not in evidence

Failure 103 — `SEC\_10K\_multi\_0108\_02` · HIGH confidence

CROSS-REFERENCE MISS

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Cross-Reference
- **Question:** Identify the two Co-CEOs of CIB and describe their respective share holding requirements.
- ✓ **Ground truth:** Douglas Petno is required to hold 50% of the net issued shares received as a result of PSUs vesting on March 25, 2026, while Troy Rohrbaugh is required to hold 50,000 shares.
- ✗ **Model answered:** The provided evidence does not identify the two Co-CEOs of CIB or state their respective share holding requirements.

It only describes general employee share-based awards, including RSUs and PSUs, and notes that vested PSU shares for Operating Committee members may be subject to additional holding periods.

- **Doctor diagnosis:** **Dr. Default** model error
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 104 — `SEC\_10K\_multi\_neg\_1113\_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 was the specific interest rate reduction applied to the loans that received both an interest rate reduction and a term extension?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** The loans that received both an interest rate reduction and a term extension had their weighted-average contractual interest reduced by **672 basis points**.

- **Doctor diagnosis:** **Dr. No** model error
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 105 — `SEC\_10K\_multi\_neg\_0211\_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 specific date or fiscal quarter associated with the column where total net revenue was \$185,581?
- **✓ Ground truth:** UNANSWERABLE
- **✗ Model answered:** The \$185,581 total net revenue figure is associated with the full year ended **December 31, 2025**.
- **Doctor diagnosis:** **Dr. No** model error
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 106 — `SEC\_10K\_multi\_neg\_1095\_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 was the total value of loans originated in 2020?
- **✓ Ground truth:** UNANSWERABLE
- **✗ Model answered:** The total value of loans originated in 2020 was **\$52.397 billion** (\$52,397 million).
- **Doctor diagnosis:** **Dr. No** model error
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer; Fabricated number: '\$52.397' not in evidence

Failure 107 — `SEC\_10K\_multi\_0997\_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:** Considering the rule for amortizing gains or losses, what was the threshold value in millions of dollars that would have triggered amortization at the beginning of 2024, and what was the primary reason for the net loss recorded for that year?
- **✓ Ground truth:** The threshold value at the beginning of 2024 was \$2,220.1 million. The net loss for 2024 was attributable to lower than expected returns on plan assets.
- **✗ Model answered:** The amortization threshold at the beginning of 2024 was **\$726 million** - 10% of the greater beginning-of-year amount, which was plan assets of **\$7,260 million**.

The primary driver of the actuarial result for 2024 was a **change in the weighted-average discount rate assumption**; specifically, an increase in the discount rate produced an actuarial gain for the U.S. qualified pension plan.

- **Doctor diagnosis:** **Dr. Default** model error
- **Verifier verdict:** model error (confidence: high)

- **Validator flags:** Fabricated number: '\$726' not in evidence

Failure 108 — `t2\_SEC\_10K\_multi\_1478\_00` · HIGH confidence

BAD QUESTION

- **Category:** Exception
- **Question:** The OECD's Pillar Two model rules, which establish a global per-country minimum tax of 15%, are required by an EU directive to become effective for fiscal years starting on or after December 31, 2023. What is the exception to this requirement?
- ✓ **Ground truth:** The EU directive requiring member states to incorporate Pillar Two model rules into their domestic laws has certain limited exceptions.
- ✗ **Model answered:** The evidence does not specify the exceptions; it only states that the EU directive applies "with certain limited exceptions."
- **Doctor diagnosis:** **Dr. Default** bad test
- **Verifier verdict:** bad test (confidence: high)
- **Validator flags:** (none)

Failure 109 — `SEC\_10K\_multi\_neg\_0047\_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 specific event or change in circumstances led to the reduction of the valuation allowance from \$1,610 million to \$768 million between January 26, 2025, and January 25, 2026?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** The reduction was driven by a reassessment of certain state deferred tax assets. Based on recent jurisdictional taxable income and expected future earnings, management concluded those state deferred tax assets were more likely than not realizable, leading to a \$711 million release of the valuation allowance.
- **Doctor diagnosis:** **Dr. No** model error
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 110 — `SEC\_10K\_multi\_neg\_0526\_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:** How many days must a wholesale loan be past due before it is classified as nonperforming?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** There is no fixed minimum for wholesale loans: they may be classified as nonperforming before they are 90 days past due if full payment of principal and interest is not expected. Otherwise, loans are generally considered nonperforming when principal and interest have been in default for 90 days or more, unless well-secured and in collection.
- **Doctor diagnosis:** **Dr. No** model error
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 111 — `t2\_SEC\_10K\_multi\_0873\_00` · HIGH confidence

BAD QUESTION

- **Category:** Cross-Reference
- **Question:** What is the combined total fair value of trading assets and liabilities for both Foreign exchange and Equity?
- ✓ **Ground truth:** The combined total fair value is \$348,507, which is the sum of the Foreign exchange total fair value (\$262,920) and the Equity total fair value (\$85,220).
- ✗ **Model answered:** For December 31, 2025, combining the total fair value for Foreign exchange and Equity across trading assets and trading liabilities:
 - Foreign exchange: \$19,355 million assets + \$14,833 million liabilities
 - Equity: \$5,867 million assets + \$14,806 million liabilities

Combined total = **\$54,861 million**.

- **Doctor diagnosis:** **Dr. Default** bad test
- **Verifier verdict:** bad test (confidence: high)
- **Validator flags:** Fabricated number: '\$54,861' not in evidence

Failure 112 — `SEC\_10K\_multi\_neg\_0047\_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:** For the fiscal year ended January 28, 2024, what was the total dollar amount of stock-based compensation that led to the tax benefit of \$741 million?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** For the fiscal year ended January 28, 2024, stock-based compensation expense was **\$3.549 billion**, associated with the reported **\$741 million tax benefit**.
- **Doctor diagnosis:** **Dr. No** model error
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 113 — `SEC\_10K\_multi\_neg\_0605\_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 maximum allowable holding period for calculating Risk Management VaR?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** Risk Management VaR is calculated assuming a **one-day holding period**.
- **Doctor diagnosis:** **Dr. No** model error
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 114 — `SEC\_10K\_multi\_0997\_00` · HIGH confidence

BAD QUESTION

- **Category:** Cross-Reference

- **Question:** For the year ended December 31, 2025, what was the minimum threshold in millions of dollars for recognizing amortization of net gains or losses, and was the net gain for that year attributable to higher or lower than expected returns on plan assets?
- ✓ **Ground truth:** The minimum threshold for recognizing amortization was \$2,360.3 million, and the net gain for 2025 was attributable to higher than expected returns on plan assets.
- ✗ **Model answered:** The minimum threshold was **\$2,220.1 million** (10% of the greater of the beginning-of-year projected benefit obligation of \$14,459 million or plan assets of \$22,201 million).

For 2025, the net gain was attributable to **higher than expected returns on plan assets**.

- **Doctor diagnosis:** **Dr. Default** bad test
- **Verifier verdict:** bad test (confidence: high)
- **Validator flags:** Fabricated number: '\$2,220.1' not in evidence

Failure 115 — `SEC\_10K\_multi\_0049\_02` · 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:** Considering the fiscal year ended January 28, 2024, what was the sum of the revenue from Taiwan and the stock-based compensation expense?
- ✓ **Ground truth:** For the fiscal year ended January 28, 2024, the sum of revenue from Taiwan (\$14,912 million) and stock-based compensation expense (\$3,549 million) was \$11,363 million.
- ✗ **Model answered:** For the fiscal year ended January 28, 2024:
 - Taiwan revenue: \$14,912 million
 - Stock-based compensation expense: \$3,549 million

Sum: **\$18,461 million** (or **\$18.461 billion**)

- **Doctor diagnosis:** **Dr. Default** model error
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Fabricated number: '\$18,461' not in evidence

Failure 116 — `SEC\_10K\_multi\_1137\_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:** By how much did the combined lending-related commitments for the consumer (excluding credit card) and credit card portfolio segments increase from December 31, 2024 to December 31, 2025?
- ✓ **Ground truth:** The combined lending-related commitments for the consumer (excluding credit card) and credit card portfolio segments increased by \$0.2 trillion (or \$200 billion) from December 31, 2024 to December 31, 2025.
- ✗ **Model answered:** The combined lending-related commitments increased by **\$175,198 million**, or approximately **\$175.2 billion**.

Calculation:

- 2025: \$43,587 million + \$1,177,766 million = **\$1,221,353 million**
- 2024: \$44,844 million + \$1,001,311 million = **\$1,046,155 million**
- Increase: **\$1,221,353 million – \$1,046,155 million = \$175,198 million**
- **Doctor diagnosis:** **Dr. Default** model error
- **Verifier verdict:** model error (confidence: high)

- **Validator flags:** Fabricated number: '\$175,198' not in evidence

Failure 117 — `SEC\_10K\_multi\_0691\_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 value of Commercial nonagency mortgage-backed securities, in millions?
- ✓ **Ground truth:** 4,065
- ✗ **Model answered:** At December 31, 2025, Commercial nonagency mortgage-backed securities totaled \$6,421 million:
- Trading assets: **\$1,937 million**
- Available-for-sale securities: **\$4,484 million**
- **Doctor diagnosis:** **Dr. Default** model error
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Fabricated number: '\$6,421' not in evidence

Failure 118 — `t2\_SEC\_10K\_multi\_0964\_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 net fair value for all noninvestment-grade reference entities with maturities of 5 years or less?
- ✓ **Ground truth:** The total net fair value for noninvestment-grade entities with maturities of 5 years or less is \$(135,083) million, calculated by summing the net fair value for maturities <1 year (\$(43,863) million) and 1-5 years (\$(91,220) million).
- ✗ **Model answered:** The evidence does not provide net fair value broken out by maturity.

For December 31, 2025, the total net fair value for all noninvestment-grade reference entities is **\$1.354 billion**, but this includes all maturities, including those over 5 years.

- **Doctor diagnosis:** **Dr. Default** model error
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Fabricated number: '\$1.354' not in evidence

Failure 119 — `SEC\_10K\_multi\_neg\_1432\_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 value of noninterest-bearing liabilities?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** The total value of noninterest-bearing liabilities is **\$670,426 million**.
- **Doctor diagnosis:** **Dr. No** model error
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 120 — `SEC\_10K\_multi\_neg\_1435\_05` · HIGH confidence

BAD QUESTION

- **Category:** Negative (hallucination traps)
- **Question:** What is the specific dollar amount of noninterest-bearing liabilities that were funding interest-earning assets in 2023?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:**
- **Doctor diagnosis:** **Dr. No** bad test
- **Verifier verdict:** bad test (confidence: high)
- **Validator flags:** No answer: model returned an empty response

Failure 121 — `SEC\_10K\_multi\_neg\_0211\_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 was the income tax expense for the period where pre-provision profit was \$75,194?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** The income tax expense was **\$16.322 billion** for the year ended December 31, 2023, when pre-provision profit was **\$75.194 billion**.
- **Doctor diagnosis:** **Dr. No** model error
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 122 — `SEC\_10K\_multi\_neg\_0452\_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 excise tax rate that is excluded from the common share repurchase calculations?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** The excise tax rate excluded from the common share repurchase calculations is **1%**.
- **Doctor diagnosis:** **Dr. No** model error
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 123 — `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:** Cross-Reference
- **Question:** Considering the total retained loans for the current period, what percentage is comprised of loans from the top two geographic regions combined?
- ✓ **Ground truth:** The top two geographic regions, California (\$117,500) and New York (\$46,378), combine for a total of \$163,878. This represents 54.0% of the total retained loans for the current period (\$303,531).

- **✗ Model answered:** For the current period (December 31, 2025) credit card retained loans, the top two geographic regions are:

- California: \$38,702 million
- Texas: \$26,313 million
- Combined: \$65,015 million

Total retained loans: \$247,797 million

Percentage = $\$65,015 \div \$247,797 \times 100 = 26.2\%$.

- **Doctor diagnosis:** **Dr. Default** model error
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Fabricated number: '100' not in evidence

Failure 124 — `SEC\_10K\_multi\_0038\_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 was the net unrealized gain on non-marketable equity securities for the period ending January 26, 2025, and what percentage of the total accounts receivable balance for that same period did the top two direct customers represent?
- **✓ Ground truth:** The net unrealized gain on non-marketable equity securities was \$995 million as of January 26, 2025, and the top two direct customers accounted for 33% of the accounts receivable balance for that same date.
- **✗ Model answered:** For the period ending January 26, 2025:
 - Net unrealized gain on non-marketable equity securities: \$757 million, calculated as \$816 million unrealized gains less \$59 million impairments and unrealized losses.
 - Top two direct customers' share of accounts receivable: 33% combined, from 17% and 16%.
- **Doctor diagnosis:** **Dr. Default** model error
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Fabricated number: '\$757' not in evidence

Failure 125 — `SEC\_10K\_multi\_1152\_00` · HIGH confidence

PRECISION ERROR

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Precision
- **Question:** What was the value of senior securities excluded from certain calculations as of December 31, 2025?
- **✓ Ground truth:** 188 million
- **✗ Model answered:** As of December 31, 2025, the carrying amount excluded for AFS securities was **\$33.6 billion**.
- **Doctor diagnosis:** **Dr. Default** model error
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 126 — `SEC\_10K\_multi\_0014\_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:** By how much does the amortized cost of Corporate debt securities exceed the combined amortized cost of all government securities (U.S. Treasury, U.S. agency, and Non-U.S. government)?
- ✓ **Ground truth:** The amortized cost of Corporate debt securities is \$65,622. The combined amortized cost for U.S. Treasury, U.S. agency, and Non-U.S. government securities is \$39,540 (\$16,150 + \$5,431 + \$17,959). Therefore, Corporate debt securities exceed the combined government securities by \$26,082.
- ✗ **Model answered:** Corporate debt securities exceed the combined amortized cost of government securities by **\$18.589 billion**.

Calculation:

- Corporate debt securities: **\$32.887 billion**
- Government securities: U.S. Treasury **\$11.095B** + U.S. agencies **\$3.163B** + Non-U.S. government **\$0.040B** = **\$14.298B**
- Difference: **\$32.887B – \$14.298B = \$18.589B**
- **Doctor diagnosis:** **Dr. Default** model error
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Fabricated number: '095B' not in evidence

Failure 127 — `SEC\_10K\_multi\_neg\_0005\_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 were the total net sales for the Services category in 2025?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** The Services category had total net sales of **\$109.158 billion** in 2025.
- **Doctor diagnosis:** **Dr. No** model error
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer; Fabricated number: '\$109.158' not in evidence

Failure 128 — `SEC\_10K\_multi\_neg\_1273\_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 vesting period for the Restricted Stock Units (RSUs) that are treated as participating securities?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** The RSUs treated as participating securities generally vest **50% after two years and 50% after three years**.
- **Doctor diagnosis:** **Dr. No** model error
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 129 — `SEC\_10K\_multi\_1096\_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:** For the year ended December 31, 2024, what was the total value, in millions, of investment-grade term loans originated in 2023?
- ✓ **Ground truth:** \$17,039
- ✗ **Model answered:** \$33,158 million.

Calculation: \$9,834 million + \$6,285 million + \$17,039 million = \$33,158 million.

- **Doctor diagnosis:** **Dr. Default** model error
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Fabricated number: '\$33,158' not in evidence

Failure 130 — `SEC\_10K\_multi\_neg\_0020\_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 was the total amount of gross unrecognized tax benefits for Apple as of September 28, 2024?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** Apple's total gross unrecognized tax benefits as of September 28, 2024 were **\$22.0 billion**.
- **Doctor diagnosis:** **Dr. No** model error
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 131 — `SEC\_10K\_multi\_neg\_0972\_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 were the total deposit-related fees for the year ended December 31, 2025?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** The total deposit-related fees for the year ended December 31, 2025 were **\$6.876 billion**.
- **Doctor diagnosis:** **Dr. No** model error
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer; Fabricated number: '\$6.876' not in evidence

Failure 132 — `SEC\_10K\_multi\_0873\_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 for Foreign exchange, including both Level 1 and Level 2 assets?
- **Ground truth:** \$254,567
- **Model answered:** For Foreign exchange trading derivatives (Level 1 + Level 2):
 - Trading assets / derivative receivables: \$111m + \$181,318m = \$181,429m
 - Trading liabilities / derivative payables: \$118m + \$176,521m = \$176,639m

Combined total: **\$358,068 million.**

- **Doctor diagnosis:** **Dr. Default** model error
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Fabricated number: '\$111m' not in evidence

Failure 133 — `SEC\_10K\_multi\_0028\_00` · HIGH confidence

BAD QUESTION

- **Category:** Cross-Reference
- **Question:** What was the total value of shares removed from the market through both share repurchases and restricted stock unit withholding during fiscal year 2026?
- **Ground truth:** The total value was \$9.4 billion, which includes \$1.5 billion from share repurchases between January 26, 2026, and February 20, 2026, and \$7.9 billion from withholding shares for employee tax obligations during fiscal year 2026.
- **Model answered:** The total value was **\$48.3 billion.**

Calculation: **\$40.4 billion** in share repurchases + **\$7.9 billion** in RSU share withholding = **\$48.3 billion.**

- **Doctor diagnosis:** **Dr. Default** bad test
- **Verifier verdict:** bad test (confidence: high)
- **Validator flags:** Fabricated number: '\$48.3' not in evidence

Failure 134 — `SEC\_10K\_multi\_0546\_00` · HIGH confidence

CROSS-REFERENCE MISS

- **Severity:** MEDIUM — uncharacterized error — review individually
- **Category:** Cross-Reference
- **Question:** What is the combined value from the second column for the Technology, Media & Telecommunications sector and the third column for the Industrials sector?
- **Ground truth:** The combined value is 75,933, calculated by adding 45,021 (from the second column for Technology, Media & Telecommunications) and 30,912 (from the third column for Industrials).
- **Model answered:** \$26,005 million + \$1,146 million = \$27,151 million.
- **Doctor diagnosis:** **Dr. Default** model error
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** (none)

Failure 135 — `SEC\_10K\_multi\_neg\_1273\_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 total amount, in dollars, that the Board of Directors has authorized for the common share repurchase program?

- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** \$50 billion.
- **Doctor diagnosis:** **Dr. No** model error
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 136 — `SEC\_10K\_multi\_neg\_0493\_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 was the specific long-term issuer rating for JPMorgan Chase Bank, N.A. that Moody's affirmed on May 19, 2025?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** Moody's affirmed JPMorgan Chase Bank, N.A.'s long-term issuer rating of **Aa2** on May 19, 2025.
- **Doctor diagnosis:** **Dr. No** model error
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

Failure 137 — `SEC\_10K\_multi\_neg\_1111\_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:** For single modifications involving a term extension, what was the maximum number of months by which any single loan was extended?
- ✓ **Ground truth:** UNANSWERABLE
- ✗ **Model answered:** 28 months - for "Other" loans with single term-extension modifications in 2024, measured on a weighted-average basis.
- **Doctor diagnosis:** **Dr. No** model error
- **Verifier verdict:** model error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

EU AI Act — Conformity Evidence Statement

This report documents accuracy and robustness testing of the model under test against a representative, purpose-generated question set, with independent verification of each verdict. The methodology and results are intended to serve as **accuracy and robustness evidence in support of Article 15 (Accuracy, Robustness and Cybersecurity)** of Regulation (EU) 2024/1689 (the EU AI Act), and as a testing-and-validation record suitable for inclusion in the technical documentation under Article 11 and Annex IV.

This document is testing evidence only. It does not by itself constitute, certify, or confer conformity with the EU AI Act, and does not replace the conformity assessment, risk-management, human-oversight, or other obligations that rest with the system's provider or deployer.