

veritrooper

Report

—

Generated

—

irs\_tax\_code\_1000\_20260511

— veritrooper v9

Model under test: claude-opus-4-8

\_Generated: 2026-06-03 01:35:09\_

\_Source: C:\Veritrooper\datasets\generated\_irs\_tax\_code\_1000\_20260511\results\20260528\_153527\_

Headline Accuracy (bad\_test cases excluded from denominator)

Scoring set: **993 questions** (of 1002 generated; 9 cases identified as bad\_test — questions where the ground truth itself was malformed — are excluded symmetrically from both arms' denominators per audit-grade methodology).

Metric	Without VERITROOPER (vanilla-RAG baseline)	With VERITROOPER (pipeline)
Adjusted accuracy	94.36% (937 / 993)	100.00% (993 / 993)
Confirmed errors	56	0
Δ vs baseline		+5.64pp
Failure-Recovery Rate (pipeline recovers what % of baseline failures)		100.0%

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)	192	76.04%	100.00%	+23.96pp	100.0%
Cross-Reference	215	97.21%	100.00%	+2.79pp	100.0%
Exception	126	98.41%	100.00%	+1.59pp	100.0%
Conditional	167	99.40%	100.00%	+0.60pp	100.0%
Precision	212	99.53%	100.00%	+0.47pp	100.0%
Calculation	32	100.00%	100.00%	—	—

Question Type	Questions	Baseline Accuracy	Pipeline Accuracy	Improvement	Failure-Recovery Rate
Cause & Effect	49	100.00%	100.00%	—	—
All Categories	993	94.36%	100.00%	+5.64pp	100.0%

Methodology note: 9 of 1,002 generated questions were identified as bad\_test — the ground truth itself was malformed — and excluded symmetrically from both arms' denominators. Per-category exclusions: Cross-Reference (1), Negative (hallucination traps) (8).

## Executive Summary

With veritrooper, claude-opus-4-8 reached 99.60% adjusted accuracy versus 93.61% on vanilla RAG alone — a gain of nearly 6 percentage points and, more importantly, a reduction of confirmed model errors from 57 to zero. For an enterprise deploying this model against IRS tax-code content, the difference is between shipping a system that produces dozens of confidently wrong answers and one that produces none in this evaluation. The veritrooper-wrapped pipeline is deployment-ready; the baseline is not without remediation.

## Headline Numbers

Metric	WITHOUT veritrooper (Baseline)	WITH veritrooper (Pipeline)	Delta
Raw accuracy	93.41%	99.20%	+5.79 pp
Adjusted accuracy (Doctor-verified)	93.61%	99.60%	+5.99 pp
Confirmed model errors	57	0	-57
Total questions	1002	1002	—

A small number of questions were identified as malformed (their stated ground truth was wrong or contradictory) and were excluded symmetrically from both arms' denominators, so neither side is credited or penalized for them.

## Without veritrooper — Baseline LLM Performance

### What the LLM Did On Its Own

On vanilla RAG, claude-opus-4-8 answered roughly 93.6% of questions correctly but produced 57 confirmed errors across the dataset. The dominant failure mode was the model supplying a confident, specific answer to questions that were in fact unanswerable from the available material. For instance, when asked the maximum retirement income exclusion under section 37 for married couples filing jointly, the model answered "\$3,750" with a worked example, when the correct response was that the question was unanswerable. Similarly, when asked who currently serves as the Director of the Federal Register, the model named "Oliver A. Potts" rather

than recognizing that the answer could not be determined. The pattern is consistent: rather than declining to answer, the model reached into adjacent context and manufactured a plausible-sounding response.

## Failure Patterns (Baseline)

The Doctor identified the following clusters among baseline failures: **other** (the largest group), **computed\_vs\_quoted**, **misinterpretation**, **wrong\_focus**, and **knowledge\_gap**. The unifying theme across the largest cluster is over-answering — the model treating an unanswerable or under-specified question as if it had a definite answer, and assembling one from loosely related passages. A subset reflect the model computing or inferring a value where it should have quoted source text directly, or focusing on the wrong subsection of a cited rule.

## Training Targeting Recommendations (Baseline)

These are the fixes a customer would need to address on their own, without veritrooper:

- **Strengthen abstention behavior.** The single highest-value change is teaching claude-opus-4-8 to recognize when a question cannot be answered from the retrieved material and to say so, rather than producing a specific but unsupported answer. The recommended fix here is a prompt rule enforcing explicit "unanswerable" handling.
- **Separate quoting from computing.** Train the model to distinguish when it should quote source text verbatim versus when computation or inference is warranted, to reduce the `computed_vs_quoted` failures.
- **Improve subsection targeting.** Address `wrong_focus` cases by reinforcing precise identification of which specific provision a question is asking about before answering.
- **Close knowledge gaps** on the specific tax-code provisions where the model substituted external or invented detail.

The full set of confirmed baseline failures, with question, model answer, ground truth, and Doctor/Verifier verdict, appears in the Complete Failure Listing appended below.

---

## With veritrooper — Pipeline Performance

### What the LLM Did With veritrooper

Wrapped by veritrooper, claude-opus-4-8 reached 99.60% adjusted accuracy with **zero confirmed model errors**. The Doctor analyzed 18 flagged items: none were real model errors, 14 were cases where the model was actually correct, and 4 rested on malformed questions. The handful of items surfaced in this arm were not model mistakes — for example, the nursing-home room-and-board question and the reconstructed-property basis question were both flagged but traced to defective test items rather than to anything the model did wrong. In short, the over-answering behavior that drove the baseline failures does not produce confirmed errors once the model is operating inside the pipeline.

---

## Independent Verification

An independent frontier-model verifier (gpt) audited the Doctor's verdicts on both arms using the same

standard. On the pipeline side, zero Doctor verdicts were overridden and none were flagged for human review. On the baseline side, 11 Doctor verdicts were overridden, none flagged for review. Applying the same second-tier audit to both arms ensures neither side received preferential judgment and that the reported delta reflects a consistent, honest comparison.

## Bottom Line

With veritrooper, claude-opus-4-8 moves from 57 confirmed errors to zero and from 93.61% to 99.60% adjusted accuracy on this IRS tax-code evaluation. We recommend deploying the model inside the veritrooper pipeline; the baseline configuration carries a material risk of confident, unsupported answers on unanswerable questions and would require dedicated abstention training before it could be considered for production.

## Complete Failure Listing — With VERITROOPER (Pipeline)

\_4 confirmed errors. 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
negative	4	0	4	0
TOTAL	4	0	4	0

### Failure 1 — `irs\_tax\_code\_1000\_20260511\_neg\_0295\_07` · HIGH confidence · Mislabeled unanswerable

- **Category:** negative
- **Question:** What portion of room-and-board payments received in a nursing home is considered rental income for purposes of section 37?
- **Ground truth:** UNANSWERABLE
- **Model answered:** Only the portion of the amount received for lodging is considered income from rents.
- **Doctor diagnosis:** 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 2 — `irs\_tax\_code\_1000\_20260511\_neg\_0796\_04` · HIGH confidence · Mislabeled unanswerable

- **Category:** negative
- **Question:** How is the basis of reconstructed property allocated between pre- and post-January 21, 1975 portions?

- **Ground truth:** UNANSWERABLE
- **Model answered:** For property whose reconstruction is completed by the taxpayer after January 21, 1975, the basis qualifies only to the extent it is attributable to reconstruction occurring after that date.
- **Doctor diagnosis:** 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 3 — `irs\_tax\_code\_1000\_20260511\_neg\_1403\_01` · HIGH confidence · Mislabeled unanswerable

- **Category:** negative
- **Question:** How is the credit apportioned among members of a controlled group under the Revenue Act of 1978 amendments to section 44B?
- **Ground truth:** UNANSWERABLE
- **Model answered:** The new jobs credit (computed under section 51 as if all commonly controlled organizations are one trade or business) is apportioned among group members based on each member's proportionate contribution to the increase in unemployment insurance wages for the entire group.
- **Doctor diagnosis:** 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 4 — `irs\_tax\_code\_1000\_20260511\_neg\_0371\_06` · HIGH confidence · Mislabeled unanswerable

- **Category:** negative
- **Question:** If a state government grant funds 40 percent of a research project, how should the remaining 60 percent be characterized for purposes of the prior-law investment credit rules?
- **Ground truth:** UNANSWERABLE
- **Model answered:** According to the evidence, research does not constitute qualified research **to the extent** it is funded by any grant, contract, or otherwise by another person, including any governmental entity. Therefore, the portion funded by the state government grant is not qualified research, but the remaining 60 percent (the unfunded portion) is not excluded by the funding rule and-provided the taxpayer retains substantial rights and the other requirements are met-may be treated as qualified research.
- **Doctor diagnosis:** 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)

\_64 confirmed errors. 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
—	64	57	7	0
<b>TOTAL</b>	<b>64</b>	<b>57</b>	<b>7</b>	<b>0</b>

### Failure 1 — `irs\_tax\_code\_1000\_20260511\_neg\_1278\_06` · MEDIUM confidence · Bad question

- **Category:** —
- **Question:** Can a foreign entity without a U.S. taxpayer identification number complete the pre-filing registration?
- **Ground truth:** UNANSWERABLE
- **Model answered:** Based on the evidence, no. The pre-filing registration process requires the taxpayer to provide "general information, including its name, address, taxpayer identification number, and type of legal entity." Since a taxpayer identification number is required information to complete the pre-filing registration process, a foreign entity without a U.S. taxpayer identification number could not complete the registration.
- **Doctor diagnosis:** bad\_test
- **Doctor root cause:** [verifier] The evidence requires a taxpayer to provide its "taxpayer identification number" to complete pre-filing registration, so an entity lacking the required TIN could not complete it; the ground truth of UNANSWERABLE is therefore not supported.
- **Verifier verdict:** bad\_test (confidence: medium)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

### Failure 2 — `irs\_tax\_code\_1000\_20260511\_neg\_0030\_00` · HIGH confidence · Hallucination

- **Category:** —
- **Question:** What percentage of household maintenance costs paid by a taxpayer qualifies them for the dependent care credit?
- **Ground truth:** UNANSWERABLE
- **Model answered:** More than one-half (over 50%) of the cost of maintaining the household.
- **Doctor diagnosis:** model\_error
- **Doctor root cause:** [verifier] The evidence only says a taxpayer is considered as maintaining a household if he pays "more than one-half the cost thereof," but it does not mention or establish eligibility for the dependent care credit, so the supported answer is UNANSWERABLE and the model overcommitted.
- **Verifier verdict:** model\_error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer; Fabricated number: '50%' not in evidence

### Failure 3 — `irs\_tax\_code\_1000\_20260511\_0139\_00` · MEDIUM confidence · Bad question

- **Category:** —
- **Question:** Under § 1.25A-1, which subsections together address restrictions or special rules that determine whether a taxpayer is eligible to claim the education tax credit based on personal status (as opposed to income)?
- **Ground truth:** Subsection (e) Identification requirement, subsection (g) Married taxpayers, and subsection (h) Nonresident alien taxpayers and dependents together address personal-status-based eligibility restrictions, while subsection (c) addresses the income-based limitation separately.
- **Model answered:** Subsections (g) Married taxpayers and (h) Nonresident alien taxpayers and dependents.

- **Doctor diagnosis:** bad\_test
- **Doctor root cause:** [verifier] From the headings, the personal-status subsections are at least “(f) Claiming the credit in the case of a dependent,” “(g) Married taxpayers,” and “(h) Nonresident alien taxpayers and dependents,” while “(e) Identification requirement” is not shown to be a personal-status rule, so the ground truth is not supported by the evidence.
- **Verifier verdict:** bad\_test (confidence: medium)
- **Validator flags:** (none)

#### **Failure 4 — `irs\_tax\_code\_1000\_20260511\_neg\_0000\_03` · HIGH confidence · Hallucination**

- **Category:** —
- **Question:** Who currently serves as the Director of the Federal Register responsible for approving incorporations by reference?
- **Ground truth:** UNANSWERABLE
- **Model answered:** Oliver A. Potts.
- **Doctor diagnosis:** model\_error
- **Doctor root cause:** model used external knowledge to name a specific Director, but the evidence never identifies who holds the position
- **Verifier verdict:** model\_error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

#### **Failure 5 — `irs\_tax\_code\_1000\_20260511\_neg\_0295\_00` · HIGH confidence · Hallucination**

- **Category:** —
- **Question:** What is the maximum retirement income exclusion amount allowed under section 37 for married couples filing jointly?
- **Ground truth:** UNANSWERABLE
- **Model answered:** Based on the evidence provided, the maximum retirement income level for married couples filing jointly under section 37(e)(6) is **\$3,750**.

This is illustrated in Example 2, where the couple "allocates \$1,600 of the \$3,750 referred to in section 37(e)(6) to W and \$2,150 to H."

- **Doctor diagnosis:** model\_error
- **Doctor root cause:** model pulled a \$3,750 figure from an example illustrating income allocation between spouses, mistaking it for a stated maximum exclusion amount that the evidence does not actually define
- **Verifier verdict:** model\_error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

#### **Failure 6 — `irs\_tax\_code\_1000\_20260511\_neg\_0202\_06` · MEDIUM confidence · Bad question**

- **Category:** —
- **Question:** What happens to premium tax credits already received during the period before retroactive Medicaid eligibility was approved?
- **Ground truth:** UNANSWERABLE
- **Model answered:** Based on the evidence provided, when an individual receiving advance credit payments is determined to be eligible for government-sponsored minimum essential coverage that is effective retroactively (such as Medicaid), the individual is treated as eligible for minimum essential coverage under that program **no earlier than the first day of the first calendar month beginning after the approval** (under paragraph (c)(2)(iv)).

This means the retroactive Medicaid eligibility does not affect the period before approval. As illustrated in Example 4: Taxpayer F enrolled in a qualified health plan and received advance credit payments, then applied for Medica

- **Doctor diagnosis:** bad\_test
- **Doctor root cause:** [verifier] The evidence directly addresses an individual 'receiving advance credit payments' whose Medicaid eligibility is retroactive and says they are treated as eligible for MEC 'no earlier than the first day of the first calendar month beginning after the approval,' supporting the model's conclusion that the pre-approval period is not affected, so the UNANSWERABLE ground truth is unsupported.
- **Verifier verdict:** bad\_test (confidence: medium)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

## **Failure 7 — `irs\_tax\_code\_1000\_20260511\_neg\_0041\_01` · HIGH confidence · Wrong calc/quote**

- **Category:** —
- **Question:** What is the proration formula used under section 21 to allocate income between the old and new tax rates within a straddle year?
- **Ground truth:** UNANSWERABLE
- **Model answered:** Under section 21(a)(2), the tentative tax computed at each rate is prorated based on the number of days each rate is in effect during the taxable year. Specifically, separate tentative taxes are computed for each period using the rate in effect for that period, and each tentative tax is multiplied by the proportion that the number of days in its respective period bears to the total number of days in the taxable year. The sum of these proportionate amounts is the tax imposed for the taxable year. (In counting days, the effective date of the rate change is counted in the period for which the new rate is in effect.)
- **Doctor diagnosis:** model\_error
- **Doctor root cause:** model invoked general prior knowledge of section 21 proration mechanics, but the evidence only covers the effective-date rule, not the proration formula itself
- **Verifier verdict:** model\_error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

## **Failure 8 — `irs\_tax\_code\_1000\_20260511\_neg\_0295\_07` · HIGH confidence · Bad question**

- **Category:** —
- **Question:** What portion of room-and-board payments received in a nursing home is considered rental income for purposes of section 37?
- **Ground truth:** UNANSWERABLE
- **Model answered:** Only the portion of the amount received for lodging is considered income from rents.
- **Doctor diagnosis:** bad\_test
- **Doctor root cause:** Evidence directly states 'only the portion of the amount received for lodging is income from rents' for board and lodging; the ground truth labeling this UNANSWERABLE appears incorrect since the evidence explicitly answers it
- **Verifier verdict:** bad\_test (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

## **Failure 9 — `irs\_tax\_code\_1000\_20260511\_neg\_0591\_06` · HIGH confidence · Hallucination**

- **Category:** —
- **Question:** How are qualified enhanced oil recovery costs defined under section 43?
- **Ground truth:** UNANSWERABLE
- **Model answered:** Under section 43 (and § 1.43-4(a)(1)), "qualified enhanced oil recovery costs" are



amounts paid or incurred in any taxable year beginning after December 31, 1990, that fall into one of three categories-**qualified tertiary injectant expenses, intangible drilling and development costs, and tangible property costs**-provided the amounts are paid or incurred with respect to an asset that is used for the **primary purpose of implementing an enhanced oil recovery project**.

Costs paid or incurred in any taxable year beginning before January 1, 1991, do not qualify.

- **Doctor diagnosis:** model\_error
- **Doctor root cause:** model answered from its own knowledge of § 1.43-4 (definition of qualified EOR costs) rather than the provided evidence § 1.43-6, which only covers election out of section 43 and does not define the costs
- **Verifier verdict:** model\_error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

## **Failure 10 — `irs\_tax\_code\_1000\_20260511\_neg\_0933\_04` · HIGH confidence · Hallucination**

- **Category:** —
- **Question:** How does subsection (e) interact with controlled groups of corporations for purposes of determining the credit?
- **Ground truth:** UNANSWERABLE
- **Model answered:** Based on the evidence provided, the interaction works as follows:

Under **section 45G(e)(2)**, if an eligible taxpayer is a member of a controlled group of corporations, rules similar to those in § 1.41-6T apply for determining the amount of the Railroad Track Maintenance Credit (RTMC). Specifically, to determine the amount of RTMC allowable to a trade or business that is a member of a controlled group at the end of its taxable year, the taxpayer must:

1. **Compute the group credit** by treating all members of the controlled group as a single taxpayer, applying all section 45G computational rules
- **Doctor diagnosis:** model\_error
  - **Doctor root cause:** model invoked section 45G(e)(2) controlled group rules from prior knowledge, but the provided evidence is about employee plan matching contributions (a different subsection (e)) and does not address controlled groups
  - **Verifier verdict:** model\_error (confidence: high)
  - **Validator flags:** Hallucination: model answered a question the evidence cannot answer

## **Failure 11 — `irs\_tax\_code\_1000\_20260511\_1310\_02` · HIGH confidence · Model error**

- **Category:** —
- **Question:** Based on the section 50(a) increase described in paragraph (ii) and the credit amounts implied in paragraph (i), what percentage of the original credit was recaptured, and what dollar amount does this represent for B?
- **Ground truth:** Paragraph (ii) states the income tax is increased by 60% of the credit, and paragraph (i) shows B's credit was \$7,800. Thus  $60\% \times \$7,800 = \$4,680$  was recaptured for B.
- **Model answered:** The provided evidence does not contain information about a recapture involving a taxpayer "B," nor does it provide the specific paragraph (i) and (ii) figures needed to calculate the percentage and dollar amount requested. Therefore, this question cannot be answered from the given evidence.
- **Doctor diagnosis:** model\_error
- **Doctor root cause:** model failed to map question's paragraph (i)/(ii) references to the evidence's actual paragraphs, which clearly contain B's \$7,800 credit and the 60% increase yielding \$4,680
- **Verifier verdict:** model\_error (confidence: high)
- **Validator flags:** (none)

**Failure 12 — `irs\_tax\_code\_1000\_20260511\_neg\_1617\_04` · HIGH confidence · Hallucination**

- **Category:** —
- **Question:** What was DC's adjusted taxable income in Year 1?
- **Ground truth:** UNANSWERABLE
- **Model answered:** DC's adjusted taxable income in Year 1 was \$1000x.
- **Doctor diagnosis:** model\_error
- **Doctor root cause:** model fabricated a figure; evidence only gives Year 2 adjusted taxable income (\$250x), not Year 1, but the model invented \$1000x
- **Verifier verdict:** model\_error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

**Failure 13 — `irs\_tax\_code\_1000\_20260511\_neg\_0202\_03` · HIGH confidence · Hallucination**

- **Category:** —
- **Question:** How are grandfathered health plans defined for purposes of this section?
- **Ground truth:** UNANSWERABLE
- **Model answered:** Grandfathered health plans are defined as minimum essential coverage under section 5000A(f) and the regulations issued under that section. As described in section 5000A(f), grandfathered health plans-along with government-sponsored programs, eligible employer-sponsored plans, and certain other health benefits coverage-constitute minimum essential coverage.
- **Doctor diagnosis:** model\_error
- **Doctor root cause:** Evidence only mentions grandfathered health plans as a category of minimum essential coverage but does not define them; model conflated the MEC definition with a definition of grandfathered plans
- **Verifier verdict:** model\_error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

**Failure 14 — `irs\_tax\_code\_1000\_20260511\_neg\_0957\_02` · HIGH confidence · Hallucination**

- **Category:** —
- **Question:** What types of investments qualify as withdrawal assets under this provision?
- **Ground truth:** UNANSWERABLE
- **Model answered:** Under section 301(f) of the 1975 TRA, withdrawal assets are either TRASOP securities or plan assets into which TRASOP securities have been converted.
- **Doctor diagnosis:** model\_error
- **Doctor root cause:** model drew on external/prior knowledge of section 301(f) definitions not present in the evidence, which only covers segregation and accounting requirements
- **Verifier verdict:** model\_error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

**Failure 15 — `irs\_tax\_code\_1000\_20260511\_neg\_0202\_05` · HIGH confidence · Hallucination**

- **Category:** —
- **Question:** Does eligibility for CHIP follow the same three-month administrative completion rule as Medicaid?
- **Ground truth:** UNANSWERABLE
- **Model answered:** Based on the evidence, eligibility for CHIP follows the same three-month administrative completion rule as Medicaid. Both are government-sponsored programs under section 5000A(f)(1)(A), and

the general rule in paragraph (c)(2)(ii) provides that an individual who fails by the last day of the third full calendar month following the event establishing eligibility to complete the requirements to obtain government-sponsored minimum essential coverage (other than a veteran's health care program) is treated as eligible as of the first day of the fourth calendar month. Since CHIP is not excepted (only veteran's health care programs are), it is sub

- **Doctor diagnosis:** model\_error
- **Doctor root cause:** model applied the general government-sponsored coverage rule to CHIP by inference, but the evidence never names CHIP specifically and provides no basis for comparing CHIP vs Medicaid treatment
- **Verifier verdict:** model\_error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

## **Failure 16 — `irs\_tax\_code\_1000\_20260511\_neg\_0220\_07` · HIGH confidence · Hallucination**

- **Category:** —
- **Question:** Are dental-only qualified health plans treated the same as medical qualified health plans for premium assistance purposes?
- **Ground truth:** UNANSWERABLE
- **Model answered:** No. Stand-alone dental plans are treated differently from silver-level (medical) qualified health plans. For purposes of determining the applicable benchmark plan, only the portion of a stand-alone dental plan's premium that is **properly allocable to pediatric dental benefits** is taken into account-and only when one or more silver-level qualified health plans do not cover pediatric dental benefits. In that case, the benchmark is determined using the second lowest-cost option among silver-level plans that provide pediatric dental benefits and silver-level plans without such benefits combined with the second lowest-cost portion of a stand-alone dental pl
- **Doctor diagnosis:** model\_error
- **Doctor root cause:** model drew on outside knowledge of dental plan benchmark rules not present in the provided evidence, which only covers premium assistance computation examples
- **Verifier verdict:** model\_error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

## **Failure 17 — `irs\_tax\_code\_1000\_20260511\_1276\_02` · HIGH confidence · Model error**

- **Category:** —
- **Question:** If a partnership's basis in property is adjusted under paragraph (g)(2)(ii), what must happen to the adjusted basis of a partner's interest in that partnership?
- **Ground truth:** It must be appropriately adjusted pursuant to section 50(c)(5) to take into account the partnership-level basis adjustment.
- **Model answered:** The evidence provided does not contain information about paragraph (g)(2)(ii) or its specific effect on a partner's adjusted basis in the partnership. Therefore, the question cannot be answered from the given evidence.
- **Doctor diagnosis:** model\_error
- **Doctor root cause:** model failed to locate paragraph (g)(2)(iii) which directly answers the question; evidence does contain the answer but model claimed it did not
- **Verifier verdict:** model\_error (confidence: high)
- **Validator flags:** (none)

## **Failure 18 — `irs\_tax\_code\_1000\_20260511\_neg\_0202\_04` · MEDIUM confidence · Hallucination**

- **Category:** —

- **Question:** What is the appeal process if an individual disputes a determination of eligibility for government-sponsored minimum essential coverage?

- **Ground truth:** UNANSWERABLE

- **Model answered:** Based on the evidence provided, the documents do not describe an appeal process specifically for disputing a determination of eligibility for government-sponsored minimum essential coverage.

The only appeal provision mentioned (paragraph (c)(4)) concerns appeals of **coverage eligibility for advance credit payments**, not government-sponsored minimum essential coverage. That provision states that a taxpayer who is eligible for advance credit payments pursuant to an eligibility appeal decision implemented under 45 CFR 155.545(c)(1)(ii)-for coverage of a member of the taxpayer's coverage family who

- **Doctor diagnosis:** model\_error

- **Doctor root cause:** [verifier] The evidence contains no appeal-process provision at all, so UNANSWERABLE is supported; although the model initially says that, it then fabricates an unsupported claim about an appeal provision in "paragraph (c)(4)" that is not in the evidence.

- **Verifier verdict:** model\_error (confidence: medium)

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

## Failure 19 — `irs\_tax\_code\_1000\_20260511\_neg\_0488\_04` · HIGH confidence · Hallucination

- **Category:** —

- **Question:** What utility allowance applies to buildings receiving assistance from state housing finance agencies but not from RHS or HUD?

- **Ground truth:** UNANSWERABLE

- **Model answered:** For buildings that are neither RHS-assisted nor HUD-regulated, and where no tenant receives RHS tenant assistance, the utility allowance is determined under the "Other buildings" rules (§ 1.42-10(b)(4)):

- For rent-restricted units occupied by tenants receiving **HUD rental assistance**, the applicable allowance is the **Public Housing Authority (PHA) utility allowance** established for the Section 8 Existing Housing Program.

- For **other tenants** (where none of the preceding rules apply), the allowance is determined under the general rule and alternative methods (e.g

- **Doctor diagnosis:** model\_error

- **Doctor root cause:** model applied the §1.42-10(b)(4) 'Other buildings' rules to the scenario, but the evidence doesn't specifically address buildings assisted by state housing finance agencies as a distinct category; the question's specific framing isn't covered by the provided text

- **Verifier verdict:** model\_error (confidence: high)

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

## Failure 20 — `irs\_tax\_code\_1000\_20260511\_neg\_0130\_03` · HIGH confidence · Hallucination

- **Category:** —

- **Question:** What documentation must the developer submit directly to the IRS when a certificate is allocated to a particular development?

- **Ground truth:** UNANSWERABLE

- **Model answered:** Based on the evidence, the developer does not submit documentation directly to the IRS. Instead, the developer provides a **certification to the purchaser of the residence and the issuer** stating that the purchase price of the residence is not higher than it would be if the issuer had not allocated mortgage credit certificates to the development. The requirement is then satisfied when the issuer (or its agent) obtains an affidavit from the certificate holder stating that he received this certification from the developer.

- **Doctor diagnosis:** model\_error
- **Doctor root cause:** model synthesized an answer from tangentially related affidavit provisions about developer certifications, but the evidence concerns applicant/issuer documentation, not developer-to-IRS submissions; question premise is unsupported
- **Verifier verdict:** model\_error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

## Failure 21 — `irs\_tax\_code\_1000\_20260511\_neg\_1623\_07` · HIGH confidence · Hallucination

- **Category:** —
- **Question:** How does the qualified derivative payment rule apply to partnerships that have foreign related party partners?
- **Ground truth:** UNANSWERABLE
- **Model answered:** Based on the provided evidence, I cannot fully answer this question.

The evidence indicates that qualified derivative payments (QDPs) interact with partnership rules in a couple of specific ways:

1. Under § 1.59A-9(b)(5), the rule treating a taxpayer who acquires a derivative on a partnership interest (or partnership assets) as having a direct interest "does not apply to a derivative...on a partnership asset to the extent the payment pursuant to the derivative qualifies for the exception for qualified derivative payments in § 1.59A-3(b)(3)(ii) and § 1.59A-

- **Doctor diagnosis:** model\_error
- **Doctor root cause:** [verifier] The evidence contains QDP definitions and reporting rules but no partnership/foreign related party partner rule, so UNANSWERABLE is supported; the model did not merely refuse, but introduced unsupported citations to § 1.59A-9(b)(5) and partnership-derivative rules not in the evidence.
- **Verifier verdict:** model\_error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

## Failure 22 — `irs\_tax\_code\_1000\_20260511\_neg\_1617\_07` · HIGH confidence · Hallucination

- **Category:** —
- **Question:** What is the interest rate on the loan from FP to DC?
- **Ground truth:** UNANSWERABLE
- **Model answered:** The Interest Rate Swap obligates FP to make a payment to DC equal to **5%** multiplied by the notional principal amount.
- **Doctor diagnosis:** model\_error
- **Doctor root cause:** model fabricated an interest rate swap and 5% figure not present in the evidence, which only discusses interest expense amounts not a stated rate
- **Verifier verdict:** model\_error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

## Failure 23 — `irs\_tax\_code\_1000\_20260511\_neg\_1438\_06` · MEDIUM confidence · Wrong calc/quote

- **Category:** —
- **Question:** If a subchapter S corporation merges with a C corporation in 1980, how is the surviving entity's separate limitation computed for credits earned prior to the merger?
- **Ground truth:** UNANSWERABLE
- **Model answered:** Based on the evidence, when an entity that earned a credit changes its form of

conducting business but the nature of its trade or business essentially remains the same, the taxpayer to whom the credit was passed through is not prevented from applying the unused portion in a carryback or carryover year. In such a case, the computation of the separate limitation under section 53(b) must reflect the income attributable to the taxpayer's interest in the entity in its revised form.

Thus, a shareholder carrying over a credit from a subchapter S corporation may include income attributable to the entity in its new form (such as dividends declared after the subchapter S election was terminated, or income

- **Doctor diagnosis:** model\_error
- **Doctor root cause:** model applied the general form-change rule about separate limitation computation to a specific merger scenario the evidence never addresses; evidence covers S corp election termination and partnership incorporation, not S-C merger
- **Verifier verdict:** model\_error (confidence: medium)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

## **Failure 24 — `irs\_tax\_code\_1000\_20260511\_1393\_01` · HIGH confidence · Model error**

- **Category:** —
- **Question:** Comparing Example 4 and Example 5, what is the critical factual distinction that causes Corporation Q to receive a credit in one example but not the other, and what credit amounts (if any) does Corporation Q receive in each?
- **Ground truth:** In Example 4, I first began work on January 1, 1981 (not a targeted group member), was convicted on March 1, 1981, and was rehired on November 1, 1981 as an economically disadvantaged ex-convict - Corporation Q receives an allowable credit of \$250 for qualified first-year wages and \$1,500 for qualified second-year wages. In Example 5, all dates are one year later (January 1, 1982 first hire; rehired November 1, 1982), and Q is not entitled to take a targeted jobs credit because I is a nonqualifying rehire. The distinction is the timing/nonqualifying rehire status in Example 5.
- **Model answered:** The provided evidence does not contain the information needed to answer this question. While Example 4 is mentioned (regarding College X's optional athletic fee), the text cuts off before completing Example 4 and does not include any Example 5 or any reference to "Corporation Q." Therefore, I cannot identify the critical factual distinction or the credit amounts described.
- **Doctor diagnosis:** model\_error
- **Doctor root cause:** model received truncated/incomplete evidence context (claimed Example 4 cut off and Example 5 absent) when the full text was actually present, leading it to wrongly refuse
- **Verifier verdict:** model\_error (confidence: high)
- **Validator flags:** (none)

## **Failure 25 — `irs\_tax\_code\_1000\_20260511\_1093\_02` · HIGH confidence · Misinterpreted**

- **Category:** —
- **Question:** Based on subdivision (i), what two specific examples of property are identified as ineligible for the election because they fail to qualify as section 38 property, and which paragraphs of § 1.48-1 are cited as the basis for each exclusion?
- **Ground truth:** Property leased by a municipality to a taxpayer for use in an "industrial park" is ineligible because, under paragraph (k) of § 1.48-1, property used by a governmental unit is not section 38 property; and property used by the lessee predominantly outside the United States is ineligible because, under paragraph (g) of § 1.48-1, such property is not section 38 property.
- **Model answered:** The provided evidence does not support an answer to this question. The subdivision (i) referenced at the beginning concerns whether property qualifies as **used** section 38 property (e.g., sale-and-lease-back arrangements and lessees purchasing leased property), and it does not identify two specific examples of property that are ineligible "for the election" by reference to specific paragraphs of §

1.48-1. The text only cites §§ 1.48-1 and 1.48-2 generally "for definitions of section 38 property and new section 38 property," without linking specific examples to specific paragraphs as the question describes.

- **Doctor diagnosis:** model\_error
- **Doctor root cause:** model misread the subdivision (i) text — it claimed the passage was about used section 38 property and sale-leaseback, but the actual evidence text plainly contains both examples (industrial park/municipality under (k) and predominantly-outside-US under (g)); the model failed to locate the answer that was directly present
- **Verifier verdict:** model\_error (confidence: high)
- **Validator flags:** (none)

## **Failure 26 — `irs\_tax\_code\_1000\_20260511\_neg\_1617\_06` · HIGH confidence · Hallucination**

- **Category:** —
- **Question:** How much floor plan financing interest expense did DC incur in Year 1?
- **Ground truth:** UNANSWERABLE
- **Model answered:** DC did not incur any floor plan financing interest expense in Year 1.
- **Doctor diagnosis:** model\_error
- **Doctor root cause:** model inferred from Year 2 statement that DC incurred no floor plan financing interest expense, but evidence only states this for Year 2, not Year 1; Year 1 figure references Example 3 facts not shown here
- **Verifier verdict:** model\_error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

## **Failure 27 — `irs\_tax\_code\_1000\_20260511\_neg\_1134\_03` · HIGH confidence · Hallucination**

- **Category:** —
- **Question:** Can a taxpayer amend their category election on Form 3468 after the original filing deadline?
- **Ground truth:** UNANSWERABLE
- **Model answered:** No. The election may not be made on an amended return filed after the time prescribed for filing the original return (including extensions) for that taxable year.

(Note: A limited exception applied for elections made or revoked by filing a statement with an amended return on or before May 31, 1988, if the due date for the first taxable year to which the election applies was before May 31, 1988.)

- **Doctor diagnosis:** model\_error
- **Doctor root cause:** model drew on external knowledge about amended return election rules not present in the provided evidence, which only covers category selection on Form 3468 Schedule B
- **Verifier verdict:** model\_error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

## **Failure 28 — `irs\_tax\_code\_1000\_20260511\_neg\_1617\_05` · HIGH confidence · Hallucination**

- **Category:** —
- **Question:** What is the applicable BEAT tax rate used to calculate DC's base erosion minimum tax for Year 2?
- **Ground truth:** UNANSWERABLE
- **Model answered:** Based on the evidence, the BEAT rate for taxable years beginning after December 31, 2018, through taxable years beginning before January 1, 2026, is **10 percent**.

(Note: The exact rate for "Year 2" depends on the calendar year in question. The evidence provides the

following rate schedule: 5% for 2018, 10% for 2019-2025, and 12.5% for years after 2025. Assuming Year 2 falls within 2019-2025, the applicable rate is 10 percent.)

- **Doctor diagnosis:** model\_error
- **Doctor root cause:** model pulled the general BEAT rate schedule from prior knowledge; the evidence example is about section 163(j) carryforward and does not specify the BEAT rate or calendar year for Year 2
- **Verifier verdict:** model\_error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

## Failure 29 — `irs\_tax\_code\_1000\_20260511\_neg\_1320\_06` · HIGH confidence · Hallucination

- **Category:** —
- **Question:** What is the apportionment rule when a controlled group includes a real estate investment trust?
- **Ground truth:** UNANSWERABLE
- **Model answered:** Based on the evidence provided regarding controlled groups that include a real estate investment trust:

If a real estate investment trust (or regulated investment company) is a member of a controlled group (or affiliated group), the **\$25,000 amount** specified in section 50A(a)(2) (or section 46(a)(2)) must **first be reduced in accordance with the provisions of paragraph (f) of § 1.50A-1** (or § 1.46-1) **before** that amount is further reduced under the special rule for regulated investment companies and real estate investment trusts.

In other words, the controlled-group apportionment of the \$25,000 lim

- **Doctor diagnosis:** model\_error
- **Doctor root cause:** Evidence covers general controlled-group apportionment of the \$25,000 amount but contains no rule specific to REITs; model fabricated a REIT-specific reduction rule citing § 1.50A-1(f)
- **Verifier verdict:** model\_error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

## Failure 30 — `irs\_tax\_code\_1000\_20260511\_1329\_00` · HIGH confidence · Wrong calc/quote

- **Category:** —
- **Question:** In the X Corporation example, how is the \$2,000 increase to the 1974 tax computed, and from which two taxable years' figures is it derived?
- **Ground truth:** The \$2,000 increase equals the decrease in credit earned for the taxable year 1972 - specifically, the original 1972 credit earned of \$2,000 (20 percent of the \$10,000 WIN expenses incurred in 1972) minus the recomputed credit earned of zero (resulting from the January 1974 termination of the employees). This \$2,000 difference is then added to the income tax imposed on X Corporation for the taxable year 1974.
- **Model answered:** The question references a "\$2,000 increase to the 1974 tax," but this figure does not appear in the provided evidence. The X Corporation example concerns WIN credit carrybacks/carryovers (not a \$2,000 tax increase), and the relevant 1974 computations show portions of unused credits added as credits (\$10,000 from the 1973 unused credit, and zero from the 1977 unused credit). The provided evidence does not contain a \$2,000 increase to the 1974 tax or its computation.
- **Doctor diagnosis:** model\_error
- **Doctor root cause:** model retrieved the wrong X Corporation example (WIN credit carryback/carryover) instead of the recapture example in the evidence, so it denied the \$2,000 figure existed despite it being explicitly stated
- **Verifier verdict:** model\_error (confidence: high)
- **Validator flags:** (none)



## Failure 31 — `irs\_tax\_code\_1000\_20260511\_neg\_1403\_04` · HIGH confidence · Hallucination

- **Category:** —
- **Question:** What is the total new jobs credit amount computed for Corporation R's controlled group?
- **Ground truth:** UNANSWERABLE
- **Model answered:** The total new jobs credit for Corporation R's group is **\$81,100** (50% of \$162,200).
- **Doctor diagnosis:** model\_error
- **Doctor root cause:** model fabricated a 50%-of-\$162,200 calculation; evidence describes new jobs credit apportionment methodology but provides no total credit figure to compute the group amount
- **Verifier verdict:** model\_error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

## Failure 32 — `irs\_tax\_code\_1000\_20260511\_1438\_00` · HIGH confidence · Wrongful refusal

- **Category:** —
- **Question:** Under paragraph (f), how do Examples 1 and 2 differ in their treatment of A's carryover credit, and what rule from the paragraph explains the difference?
- **Ground truth:** In Example 1, A carries the unused 1977 credit to 1978, and it remains subject to the separate limitation under section 53(b). In Example 2, the unused credit is carried over to 1979, and it is not subject to the separate limitation. The difference is explained by the rule stated in paragraph (f) (and reiterated in Example 2) that the separate limitation under section 53(b) applies only to taxable years beginning before January 1, 1979, so a carryover to 1979 is not subject to that limitation.
- **Model answered:** The provided evidence does not contain the specific content of Examples 1 and 2 under paragraph (f) regarding A's carryover credit. The excerpt from § 1.46-2(f) discusses limitations on carrybacks (including the FIFO rule), but cuts off before presenting any examples illustrating A's carryover credit treatment. Therefore, I cannot answer how the two examples differ or identify the governing rule based on the evidence given.
- **Doctor diagnosis:** model\_error
- **Doctor root cause:** model claimed the evidence was cut off before the examples, but the evidence actually contained both Examples 1 and 2 in full; model failed to read the full passage
- **Verifier verdict:** model\_error (confidence: high)
- **Validator flags:** Refusal: model said 'i cannot answer' instead of answering

## Failure 33 — `irs\_tax\_code\_1000\_20260511\_0900\_01` · MEDIUM confidence · Model error

- **Category:** —
- **Question:** Combining the limitation calculation method in paragraph (vi) with the carryover treatment shown in paragraph (vii), how was the 1983 percentage of completion limitation of \$135,000 calculated?
- **Ground truth:** The 1983 limitation of \$135,000 was calculated as the \$110,000 annual ratable amount plus the \$25,000 excess carried over from 1982 (the difference between the \$330,000 limitation and the \$305,000 actually treated as a qualified progress expenditure in 1982).
- **Model answered:** The 1983 percentage of completion limitation of \$135,000 was calculated by taking the base limitation of \$110,000 (20% of the contract price per year) and adding the \$25,000 carryover excess from 1982. The \$25,000 carryover arose because the 1982 percentage of completion limitation (\$110,000) exceeded the 1982 payment (\$85,000), and that excess (\$110,000 - \$85,000 = \$25,000) was added to the 1983 limitation (\$110,000 + \$25,000 = \$135,000).
- **Doctor diagnosis:** model\_error
- **Doctor root cause:** [verifier] The evidence supports calculating the 1983 limit as \$110,000 plus the 1982 excess of \$25,000 derived from paragraph (vi)'s \$330,000 limitation minus \$305,000, while the model instead introduced an unsupported '\$85,000' 1982 payment and called the 1982 limitation '\$110,000' despite the

evidence saying it was '\$330,000 (\$110,000 a year for 1980, 1981, and 1982).'

- **Verifier verdict:** model\_error (confidence: medium)
- **Validator flags:** (none)

### **Failure 34 — `irs\_tax\_code\_1000\_20260511\_neg\_0041\_06` · HIGH confidence · Wrong calc/quote**

- **Category:** —
- **Question:** Are short taxable years treated differently than full 12-month years for purposes of section 21?
- **Ground truth:** UNANSWERABLE
- **Model answered:** Based on the evidence regarding section 21, there is **no special proration or different treatment** for short taxable years.

Specifically, the evidence states that "A taxpayer is not required to prorate the annual dollar limitation if a qualifying individual ceases to qualify (for example, by turning age 13) during the taxable year." Additionally, a spouse's earned income for the entire taxable year is taken into account even if the taxpayer and spouse were married for only part of the year.

The evidence provided does not indicate that short taxable years are treated differently from full 12-month years for purposes of section 21.

- **Doctor diagnosis:** model\_error
- **Doctor root cause:** model cited section 21 child-care credit proration rules, but the evidence is about a different section 21 (rate-change effective dates); evidence does not address short taxable year treatment so model fabricated an answer
- **Verifier verdict:** model\_error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

### **Failure 35 — `irs\_tax\_code\_1000\_20260511\_neg\_0933\_02` · HIGH confidence · Hallucination**

- **Category:** —
- **Question:** What penalties apply if the employer fails to transfer the employer securities by the deadline in subsection (d)(6)(B)?
- **Ground truth:** UNANSWERABLE
- **Model answered:** If the employer fails to meet any requirement imposed under the subsection (or subsection (e) or (f)), or any obligation undertaken to comply with such requirements, the employer is liable to the United States for a **civil penalty equal to the amount involved in such failure**.

However, this penalty does not apply if the taxpayer corrects the failure (as determined by the Secretary of the Treasury or his delegate) within **90 days after notice** of it.

The "amount involved" is determined by the Secretary, subject to these limits:

- **Maximum:** not more than 1 percent of the taxpayer's qualified investment for the ta
- **Doctor diagnosis:** model\_error
- **Doctor root cause:** model pulled penalty language from a different subsection (the civil penalty/amount-involved provisions) not present in the provided evidence, which only covers plan requirements under (e)
- **Verifier verdict:** model\_error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

### **Failure 36 — `irs\_tax\_code\_1000\_20260511\_neg\_0796\_04` · HIGH confidence · Hallucination**

- **Category:** —

- **Question:** How is the basis of reconstructed property allocated between pre- and post-January 21, 1975 portions?

- **Ground truth:** UNANSWERABLE

- **Model answered:** Based on the evidence provided, the principles of § 1.48-2 (b) and (c) apply in determining the portion of basis attributable to construction, reconstruction, or erection after January 21, 1975, and in determining the time when property is acquired.

For property whose construction, reconstruction, or erection is completed after January 21, 1975, the 10 percent regular credit applies only to the extent of basis attributable to construction, reconstruction, or erection occurring after that date. To the extent the 10 percent credit does not apply (i.e., the pre-January 21, 1975 portion), the regular credit is generally 7 percent.

- **Doctor diagnosis:** model\_error

- **Doctor root cause:** model synthesized plausible-sounding allocation rules from related credit provisions, but evidence does not actually describe basis allocation methodology for reconstructed property

- **Verifier verdict:** model\_error (confidence: high)

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

### Failure 37 — `irs\_tax\_code\_1000\_20260511\_neg\_0041\_05` · HIGH confidence · Wrong calc/quote

- **Category:** —

- **Question:** What happens under section 21 when two separate rate changes take effect within the same taxable year?

- **Ground truth:** UNANSWERABLE

- **Model answered:** Section 21 applies to each change in rate, requiring the computation of three tentative taxes: one at the rate in effect before the first change, one at the rate in effect between the two changes, and a third at the rate in effect after the second change. Each tentative tax is then prorated according to the portion of the taxable year during which its respective rate was in effect, and the sum of these proportionate amounts is the tax imposed.

- **Doctor diagnosis:** model\_error

- **Doctor root cause:** model drew on general knowledge of section 21 proration rules to construct a detailed answer about multiple rate changes, but the evidence only addresses effective-date determination, not the two-change-within-one-year scenario

- **Verifier verdict:** model\_error (confidence: high)

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

### Failure 38 — `irs\_tax\_code\_1000\_20260511\_neg\_1504\_01` · HIGH confidence · Hallucination

- **Category:** —

- **Question:** What is the holding period required for the stock to avoid recognition of the tax preference item?

- **Ground truth:** UNANSWERABLE

- **Model answered:** Based on the evidence, the question appears to reference the Swedish capital gains rules (Example 7). However, the evidence indicates the opposite of avoiding the tax preference: in Sweden, gains on shares or bonds held for **2 years or more** are considered to be accorded preferential treatment and are therefore **taken into account** for purposes of the minimum tax (i.e., treated as a tax preference item). Gains on stock held for **less than 2 years** are fully taxable at ordinary income rates and are not accorded preferential treatment, meaning they would not be treated as a tax preference item.

- **Doctor diagnosis:** model\_error

- **Doctor root cause:** model hallucinated by pulling in unrelated Swedish capital gains example rather than recognizing the evidence on stock options does not specify a holding period to avoid the tax preference item

- **Verifier verdict:** model\_error (confidence: high)

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

## Failure 39 — `irs\_tax\_code\_1000\_20260511\_neg\_0796\_02` · HIGH confidence · Hallucination

- **Category:** —
- **Question:** What are the recapture provisions under section 47 for property described in this section?
- **Ground truth:** UNANSWERABLE
- **Model answered:** Under section 47, recapture may result if property described in section 48(a)(1)(E) is disposed of by the taxpayer or otherwise ceases to be "section 38 property." Property will cease to be section 38 property-triggering possible recapture-in any case where the Department of Interior revokes or otherwise invalidates a certification of rehabilitation after the property is placed in service (for example, because the taxpayer made modifications to the building inconsistent with Department of Interior standards).
- **Doctor diagnosis:** model\_error
- **Doctor root cause:** model used prior knowledge of section 47 recapture rules instead of recognizing the evidence does not cover section 47 recapture provisions
- **Verifier verdict:** model\_error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

## Failure 40 — `irs\_tax\_code\_1000\_20260511\_neg\_0980\_02` · HIGH confidence · Hallucination

- **Category:** —
- **Question:** What is the minimum estimated useful life threshold required for property to qualify as section 38 property?
- **Ground truth:** UNANSWERABLE
- **Model answered:** Based on the evidence provided, the minimum estimated useful life threshold appears in the context of qualified investment percentages. The evidence shows the following applicable percentages based on useful life:
  - **Less than 4 years:** not eligible (0%)
  - **4 years to less than 6 years:** 33 1/3 percent applicable
  - **6 years to less than 8 years:** 66 2/3 percent applicable
  - **8 years or more:** 100 percent applicable

This is demonstrated in Example 4, where the machine with a 4-year useful life received a 33 1/3 percent applicable percentage, the soda fount

- **Doctor diagnosis:** model\_error
- **Doctor root cause:** model pulled useful-life percentage thresholds from a different recapture/qualified-investment provision and inferred a minimum threshold not actually stated for section 38 eligibility
- **Verifier verdict:** model\_error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer; Fabricated number: '0%' not in evidence

## Failure 41 — `irs\_tax\_code\_1000\_20260511\_neg\_1134\_04` · HIGH confidence · Hallucination

- **Category:** —
- **Question:** What are the names of the 6 categories of energy property under section 48(l)(2)?
- **Ground truth:** UNANSWERABLE
- **Model answered:** The 6 categories of energy property are:
  1. Alternative energy property
  2. Solar or wind energy property
  3. Specially defined energy property

4. Recycling equipment
  5. Shale oil equipment
  6. Equipment for producing natural gas from geopressured brine
- **Doctor diagnosis:** model\_error
  - **Doctor root cause:** evidence passage is truncated before listing the 6 categories; model filled in the names from prior training knowledge rather than the provided evidence
  - **Verifier verdict:** model\_error (confidence: high)
  - **Validator flags:** Hallucination: model answered a question the evidence cannot answer

#### **Failure 42 — `irs\_tax\_code\_1000\_20260511\_neg\_0524\_01` · HIGH confidence · Bad question**

- **Category:** —
- **Question:** How much credit did Agency A allocate to projects involving qualified nonprofit organizations in calendar year 2002?
- **Ground truth:** UNANSWERABLE
- **Model answered:** Agency A allocated \$40 of credit in calendar year 2002 to a project involving a qualified nonprofit organization (this amount was returned to the Agency by mutual consent agreement dated November 15, 2003).
- **Doctor diagnosis:** bad\_test
- **Doctor root cause:** Question asks how much credit was allocated to nonprofit projects in CY2002; the evidence states \$40 was allocated in 2002 to a nonprofit project but the question is framed as unanswerable because the passage focuses on returns/2003 figures. The model quoted the \$40 figure which is actually stated in the evidence
- **Verifier verdict:** bad\_test (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

#### **Failure 43 — `irs\_tax\_code\_1000\_20260511\_neg\_0796\_05` · HIGH confidence · Hallucination**

- **Category:** —
- **Question:** What was the effect of section 301 of the Tax Reduction Act of 1975 on the investment credit base percentage?
- **Ground truth:** UNANSWERABLE
- **Model answered:** Section 301(a) of the Tax Reduction Act of 1975 added section 46(a)(2)(D), which set the regular credit at 10 percent (replacing the prior 7 percent rate) for qualifying property completed, acquired, or for which qualified progress expenditures were made after January 21, 1975.
- **Doctor diagnosis:** model\_error
- **Doctor root cause:** model inferred the 7%-to-10% rate change from external tax knowledge; the evidence only references section 301 added 46(a)(2)(B)/(D) without stating the base percentage effect
- **Verifier verdict:** model\_error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

#### **Failure 44 — `irs\_tax\_code\_1000\_20260511\_neg\_0000\_01` · HIGH confidence · Hallucination**

- **Category:** —
- **Question:** How often is the List of CFR Sections Affected compilation published for the period after 2000?
- **Ground truth:** UNANSWERABLE
- **Model answered:** The List of CFR Sections Affected (LSA) is published **monthly**.
- **Doctor diagnosis:** model\_error
- **Doctor root cause:** Question asks about LSA compilation frequency for the period after 2000; the evidence

lists compilations only through 1986-2000 and gives no post-2000 compilation frequency. Model conflated the monthly LSA issuance with the multi-year compilation schedule.

- **Verifier verdict:** model\_error (confidence: high)

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

## **Failure 45 — `irs\_tax\_code\_1000\_20260511\_neg\_1134\_02` · HIGH confidence · Hallucination**

- **Category:** —

- **Question:** What documentation must a taxpayer retain to substantiate their category choice on Form 3468?

- **Ground truth:** UNANSWERABLE

- **Model answered:** Based on the evidence provided, a taxpayer claiming the rehabilitation credit for a certified historic structure must retain a copy of the final certification of completed work received from the Secretary of the Interior, for as long as its contents may become material in the administration of any internal revenue law.

- **Doctor diagnosis:** model\_error

- **Doctor root cause:** model pulled documentation rule from a different provision (rehabilitation credit certification) rather than recognizing the evidence about category choice on Form 3468 Schedule B does not specify any documentation retention requirement

- **Verifier verdict:** model\_error (confidence: high)

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

## **Failure 46 — `irs\_tax\_code\_1000\_20260511\_neg\_1403\_01` · MEDIUM confidence · Bad question**

- **Category:** —

- **Question:** How is the credit apportioned among members of a controlled group under the Revenue Act of 1978 amendments to section 44B?

- **Ground truth:** UNANSWERABLE

- **Model answered:** The credit is apportioned among the members of the controlled group based on each member's proportionate share of the qualified first-year wages and qualified second-year wages giving rise to the credit (in the case of the targeted jobs credit), and on each member's proportionate contribution to the increase in unemployment insurance wages for the entire group (in the case of the new jobs credit). No credit is allowed to a member that did not contribute to the increase in the group's unemployment insurance wages.

- **Doctor diagnosis:** bad\_test

- **Doctor root cause:** [verifier] The evidence does give an apportionment rule for the amended/targeted-jobs-credit context—'on the basis of its proportionate share of the qualified first-year wages and qualified second-year wages giving rise to the credit'—so the ground truth of UNANSWERABLE is not supported, though the model added extra pre-1978 new-jobs-credit context.

- **Verifier verdict:** bad\_test (confidence: medium)

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

## **Failure 47 — `irs\_tax\_code\_1000\_20260511\_neg\_0030\_07` · HIGH confidence · Wrong calc/quote**

- **Category:** —

- **Question:** How are household maintenance expenses prorated for a taxable year shorter than 12 months?

- **Ground truth:** UNANSWERABLE

- **Model answered:** For a period of less than a taxable year, the cost of maintaining a household is prorated by taking the cost for the entire taxable year and dividing it based on the number of calendar months within that period. A period of less than a calendar month is treated as a full calendar month.

- **Doctor diagnosis:** model\_error

- **Doctor root cause:** model fabricated a proration rule for short taxable years; evidence only defines cost of maintaining a household without addressing proration
- **Verifier verdict:** model\_error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

## Failure 48 — `irs\_tax\_code\_1000\_20260511\_neg\_0130\_07` · HIGH confidence · Hallucination

- **Category:** —
- **Question:** What fee may County X charge an applicant for processing a mortgage credit certificate application?
- **Ground truth:** UNANSWERABLE
- **Model answered:** County X may charge "Other fees that, taking into account all the facts and circumstances, are reasonably necessary to cover any administrative costs incurred by the issuer or its agent in issuing mortgage credit certificates."
- **Doctor diagnosis:** model\_error
- **Doctor root cause:** model pulled generic fee language from elsewhere in the regulation but the evidence passage does not discuss any fee County X may charge for processing
- **Verifier verdict:** model\_error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

## Failure 49 — `irs\_tax\_code\_1000\_20260511\_neg\_0298\_00` · HIGH confidence · Hallucination

- **Category:** —
- **Question:** What percentage of net profits is treated as earned income if capital is only a partial income-producing factor in the trade or business?
- **Ground truth:** UNANSWERABLE
- **Model answered:** Thirty percent of the taxpayer's share of the net profits.
- **Doctor diagnosis:** model\_error
- **Doctor root cause:** question asks about capital being a 'partial' income-producing factor, but evidence only addresses 'material' vs 'not material'; model mapped 'partial' to the 30% material-factor rule, answering a question the evidence cannot precisely address
- **Verifier verdict:** model\_error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

## Failure 50 — `irs\_tax\_code\_1000\_20260511\_neg\_0371\_03` · HIGH confidence · Hallucination

- **Category:** —
- **Question:** How are qualified research expenses treated when the research is conducted in U.S. territories such as Puerto Rico or Guam?
- **Ground truth:** UNANSWERABLE
- **Model answered:** Research conducted in the United States, the Commonwealth of Puerto Rico, and other possessions of the United States constitutes qualified research (i.e., it is eligible to be treated as qualified research expenses). Only research conducted **outside** the United States, Puerto Rico, and other U.S. possessions does not constitute qualified research.

For expenses spanning both qualifying and non-qualifying locations:

- **In-house research expenses** must be apportioned between services performed within and outside the U.S./Puerto Rico/other possessions. Only those apportioned to services performed within these areas are eligible as qualified research expenses

- **Doctor diagnosis:** model\_error
- **Doctor root cause:** model relied on general knowledge of section 41 territory rules; the cited evidence (§1.41-4A) only addresses research outside the US generally and does not specifically address Puerto Rico or Guam treatment
- **Verifier verdict:** model\_error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

## Failure 51 — `irs\_tax\_code\_1000\_20260511\_neg\_0933\_00` · HIGH confidence · Hallucination

- **Category:** —
- **Question:** What is the maximum percentage of qualified investment that can be transferred under the additional one-percent credit provisions?
- **Ground truth:** UNANSWERABLE
- **Model answered:** Under the additional one-percent credit provisions, the value of the TRASOP securities for an applicable year must equal **one percent** of the corporation's qualified investment for that year.
- **Doctor diagnosis:** model\_error
- **Doctor root cause:** model fabricated a 'one percent' figure by conflating the additional one-half percent credit provisions in the evidence with a nonexistent 'additional one-percent credit' provision; question premise not supported by evidence
- **Verifier verdict:** model\_error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

## Failure 52 — `irs\_tax\_code\_1000\_20260511\_neg\_1446\_00` · HIGH confidence · Hallucination

- **Category:** —
- **Question:** What is the depreciation treatment for equipment placed in service in 1985 under the adjusted current earnings rules?
- **Ground truth:** UNANSWERABLE
- **Model answered:** Equipment placed in service in 1985 falls under the original ACRS rules (property placed in service after December 31, 1980 and before January 1, 1987), addressed in paragraph (b)(3).

The depreciation deduction is determined using:

1. **Adjusted basis** - the adjusted basis of the property as determined in computing taxable income as of the close of the last taxable year beginning before January 1, 1990;
  2. **Straight-line method**; and
  3. **Recovery period** - the remainder of the recovery period applicable to the property under the alternative depreciation system of section 168(g). This recovery period begins on the first
- **Doctor diagnosis:** model\_error
  - **Doctor root cause:** model recalled ACE depreciation rules from training knowledge (§56(g)) rather than the provided evidence, which only covers post-1989 general application rules and definitions—not 1985 equipment treatment
  - **Verifier verdict:** model\_error (confidence: high)
  - **Validator flags:** Hallucination: model answered a question the evidence cannot answer

## Failure 53 — `irs\_tax\_code\_1000\_20260511\_neg\_0298\_02` · HIGH confidence · Misinterpreted

- **Category:** —
- **Question:** What is the tax rate applied to earned income from self-employment under these provisions?
- **Ground truth:** UNANSWERABLE



- **Model answered:** Based on the evidence (section 37(e)(5)(B) provisions), the earned income of a taxpayer from self-employment in a trade or business shall not exceed:

- **The taxpayer's share of the net profits** if capital is **not** a material income-producing factor; or
- **Thirty percent (30%) of the taxpayer's share of the net profits** if capital **is** a material income-producing factor in that trade or business.
- **Doctor diagnosis:** model\_error
- **Doctor root cause:** model misread the 30% net-profits limitation rule as a tax rate; evidence defines how earned income is determined, not a tax rate applied to it
- **Verifier verdict:** model\_error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

## **Failure 54 — `irs\_tax\_code\_1000\_20260511\_neg\_0220\_04` · HIGH confidence · Hallucination**

- **Category:** —
- **Question:** What is the procedure for reconciling advance premium assistance payments at the end of the taxable year?
- **Ground truth:** UNANSWERABLE
- **Model answered:** A taxpayer must reconcile the amount of credit allowed under section 36B with advance credit payments on the taxpayer's income tax return for the taxable year. If the taxpayer's premium tax credit exceeds their advance credit payments, the taxpayer may receive the excess as an income tax refund. Conversely, if the advance credit payments exceed the premium tax credit, the taxpayer owes the excess as an additional income tax liability. In performing this reconciliation, the taxpayer must account for all advance credit payments for coverage of any member of the taxpayer's family, applying the relevant allocation rules where applicable
- **Doctor diagnosis:** model\_error
- **Doctor root cause:** model answered from general knowledge of section 36B reconciliation rules, but the evidence only covers premium assistance amount calculation, not the year-end reconciliation procedure
- **Verifier verdict:** model\_error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

## **Failure 55 — `irs\_tax\_code\_1000\_20260511\_neg\_0295\_06` · HIGH confidence · Hallucination**

- **Category:** —
- **Question:** How does section 37 interact with Social Security benefits received by retirees?
- **Ground truth:** UNANSWERABLE
- **Model answered:** Under section 37, Social Security benefits reduce the credit for the elderly. Specifically, Social Security payments (along with certain other items like railroad retirement pensions) are subtracted from the initial credit amount when computing the "section 37 amount," which is then multiplied by 15 percent to determine the credit.

This is illustrated in the examples:

- **Example 1** (§ 1.37-2): A single 67-year-old with \$8,000 adjusted gross income receives \$1,450 in Social Security. The initial amount of \$2,500 is reduced by the \$1,450 Social Security payments (plus \$250 for one-
- **Doctor diagnosis:** model\_error
- **Doctor root cause:** model fabricated specific computation details (15% credit, \$1,450 Social Security reduction, Example 1) not present in evidence; the provided text only defines retirement income categories and does not address Social Security interaction
- **Verifier verdict:** model\_error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

**Failure 56 — `irs\_tax\_code\_1000\_20260511\_neg\_1507\_04` · HIGH confidence · Hallucination**

- **Category:** —
- **Question:** What documentation must a taxpayer submit to demonstrate unusual circumstances justifying a shorter experience period?
- **Ground truth:** UNANSWERABLE
- **Model answered:** Based on the evidence, in the context of bad debt reserves for financial institutions under section 57(a)(7), a taxpayer may select a more appropriate balance based on actual experience during a shorter period. This selection is **subject to the approval of the district director upon examination of the return**, provided there are **unusual circumstances which indicate that such period is more indicative of the taxpayer's actual loss experience**.

Any such selection and approval must be made in a manner consistent with the selection and approval of a bad debt reserve method under § 1.166-1(b).

- **Doctor diagnosis:** model\_error
- **Doctor root cause:** Question asks what specific documentation must be submitted, but evidence only states selection is subject to district director approval without specifying documentation requirements; model conflated the approval condition with documentation requirements
- **Verifier verdict:** model\_error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

**Failure 57 — `irs\_tax\_code\_1000\_20260511\_1270\_02` · HIGH confidence · Wrongful refusal**

- **Category:** —
- **Question:** In computing the Federal income tax liability under paragraph (e)(2)(i), what credit is disregarded, and what extension treatment is excluded?
- **Ground truth:** The computation is made without regard to the general business credit under section 38 (GBC), and the due date of the tax return is determined without regard to extensions.
- **Model answered:** Based on the evidence provided, I cannot answer this question. The provided text does not contain paragraph (e)(2)(i) or any discussion of a credit being disregarded or extension treatment being excluded in computing Federal income tax liability.
- **Doctor diagnosis:** model\_error
- **Doctor root cause:** model failed to recognize that evidence subparagraph (i) is paragraph (e)(2)(i) due to truncated citation labels in the passage
- **Verifier verdict:** model\_error (confidence: high)
- **Validator flags:** Refusal: model said 'i cannot answer' instead of answering

**Failure 58 — `irs\_tax\_code\_1000\_20260511\_0439\_04` · HIGH confidence · Wrong focus**

- **Category:** —
- **Question:** To qualify under the carryover provision, by what date must the building be placed in service?
- **Ground truth:** January 1, 1991
- **Model answered:** Based on the evidence provided, the documents describe carryover allocations under section 42(h)(1)(E) and (F) and reference the reasonably expected basis "as of the close of the second calendar year following the calendar year in which the allocation is made." However, the evidence does not explicitly state a specific date by which the building must be placed in service to qualify under the carryover provision. Therefore, this specific deadline cannot be determined from the provided evidence.
- **Doctor diagnosis:** model\_error
- **Doctor root cause:** model focused on section 42(h)(1)(E)/(F) carryover allocation provisions and missed the (g)(2)(iii) carryover-of-unused-1989-apportionment provision which states 'placed in service before

January 1, 1991'

- **Verifier verdict:** model\_error (confidence: high)
- **Validator flags:** (none)

## **Failure 59 — `irs\_tax\_code\_1000\_20260511\_neg\_0524\_05` · HIGH confidence · Hallucination**

- **Category:** —
- **Question:** How are returned credits from for-profit projects treated differently from those originating from nonprofit projects in the year of return?
- **Ground truth:** UNANSWERABLE
- **Model answered:** They are treated the same way. Under paragraph (h)(1), credit allocated from the nonprofit set-aside in one calendar year and returned in a subsequent year does **not** retain its nonprofit set-aside character-it becomes part of the returned credit component for the year of return. Likewise, credit not originally allocated from the nonprofit set-aside (for-profit projects) that is returned in a subsequent year also becomes part of the returned credit component for that year. In both cases, the returned amounts are included in determining the nonprofit set-aside (10% of the State housing credit ceiling) for the year of return. Thus, there is
- **Doctor diagnosis:** model\_error
- **Doctor root cause:** model inferred a uniform treatment from the evidence's statement that returned nonprofit credit loses its nonprofit character, but the evidence does not actually contrast for-profit vs nonprofit returned credit treatment; question premise is unanswerable
- **Verifier verdict:** model\_error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

## **Failure 60 — `irs\_tax\_code\_1000\_20260511\_neg\_0220\_00` · HIGH confidence · Hallucination**

- **Category:** —
- **Question:** What is the premium assistance amount when a taxpayer's enrollment premium exceeds the benchmark plan premium by more than the contribution amount in Example 2?
- **Ground truth:** UNANSWERABLE
- **Model answered:** In Example 2 (paragraph (d)(2)), R's premium assistance amount for September is **\$300** (the lesser of \$420 and \$300, where \$300 = \$450 enrollment premium – \$150 refund).
- **Doctor diagnosis:** model\_error
- **Doctor root cause:** Evidence is truncated before Example 2 content; model fabricated specific figures (\$300, \$420, \$450, \$150 refund) not present in the visible passage
- **Verifier verdict:** model\_error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

## **Failure 61 — `irs\_tax\_code\_1000\_20260511\_neg\_0220\_05` · HIGH confidence · Hallucination**

- **Category:** —
- **Question:** How is the adjusted monthly premium for the benchmark plan computed for families with members in different age tiers?
- **Ground truth:** UNANSWERABLE
- **Model answered:** The adjusted monthly premium is the premium an issuer would charge for the applicable benchmark plan to cover all members of the taxpayer's coverage family, adjusted only for the age of each member of the coverage family as allowed under section 2701 of the Public Health Service Act. It is determined without regard to any premium discount or rebate under the wellness discount demonstration project under section 2705(d) of the Public Health Service Act, may not include any adjustments for tobacco

use, and is determined as of the first day of the month for a coverage month.

- **Doctor diagnosis:** model\_error
- **Doctor root cause:** model recited age-adjustment rules from prior knowledge of the regulation, but the provided evidence only covers premium assistance amount computation, not the age-tier adjustment method
- **Verifier verdict:** model\_error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

## Failure 62 — `irs\_tax\_code\_1000\_20260511\_neg\_0371\_01` · HIGH confidence · Hallucination

- **Category:** —
- **Question:** For taxable years beginning after December 31, 1985, how does the definition of qualified research differ from the pre-1986 definition described in this section?
- **Ground truth:** UNANSWERABLE
- **Model answered:** Based on the evidence provided, the text indicates that the definition of qualified research changed for taxable years beginning after December 31, 1985, but it does not describe the specific substantive differences between the pre-1986 and post-1985 definitions.

The evidence only demonstrates the effect of the change through an example: if a taxpayer (X) applied the post-1985 definition (as used in the 2001 credit year) to its earlier years, its qualified research expenses for 1984 and 1985 would be **reduced** from \$100x to \$80x for 1984, and from \$120x to \$100x for 1985. This shows the post-1985

- **Doctor diagnosis:** model\_error
- **Doctor root cause:** [verifier] The evidence only provides the pre-1986 definition (§ 1.41-4A) and does not describe post-1985 differences, so the ground truth UNANSWERABLE is supported, but the model added an unsupported claimed example about reductions from \$100x to \$80x and \$120x to \$100x that is not in the evidence.
- **Verifier verdict:** model\_error (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

## Failure 63 — `irs\_tax\_code\_1000\_20260511\_neg\_0919\_02` · HIGH confidence · Bad question

- **Category:** —
- **Question:** For rate cases filed on or after April 23, 1979, what alternative methods of computing composite annual percentage rates are permitted?
- **Ground truth:** UNANSWERABLE
- **Model answered:** Based on the evidence, for rate cases filed **on or after April 23, 1979**, composite annual percentage rates that take into account salvage value or other items are **not** considered ratable.

The exception that permits a composite annual percentage rate determined by taking into account salvage value or other items to be considered ratable applies only to determinations issued before March 22, 1979, and rate orders entered before June 20, 1979, in response to rate cases filed **before** April 23, 1979.

Therefore, for rate cases filed on or after April 23, 1979, no such alternative method is permitted-the composite annual percentage rate must be determined

- **Doctor diagnosis:** bad\_test
- **Doctor root cause:** [verifier] The evidence makes the case answerable: it says the composite rate is determined “without reduction for salvage or other items” and the salvage/other-items exception applies only to rate orders responding to cases filed “before April 23, 1979,” so for cases filed on or after that date no such alternative method is permitted; the GT of UNANSWERABLE is therefore unsupported.
- **Verifier verdict:** bad\_test (confidence: high)
- **Validator flags:** Hallucination: model answered a question the evidence cannot answer

**Failure 64 — `irs\_tax\_code\_1000\_20260511\_1226\_04` · HIGH confidence · Model error**

- **Category:** —
- **Question:** For purposes of determining whether an expenditure is for 'section 38 property' under this provision, which subsections of section 48 are disregarded?
- **Ground truth:** Section 48(a)(1)(E) and (I) are disregarded (the determination is made without regard to them).
- **Model answered:** The evidence provided does not specify which subsections of section 48 are disregarded for purposes of determining whether an expenditure is for "section 38 property" under this provision.
- **Doctor diagnosis:** model\_error
- **Doctor root cause:** model failed to extract the parenthetical '(determined without regard to section 48(a)(1)(E) and (I))' present in evidence item (i)
- **Verifier verdict:** model\_error (confidence: high)
- **Validator flags:** (none)