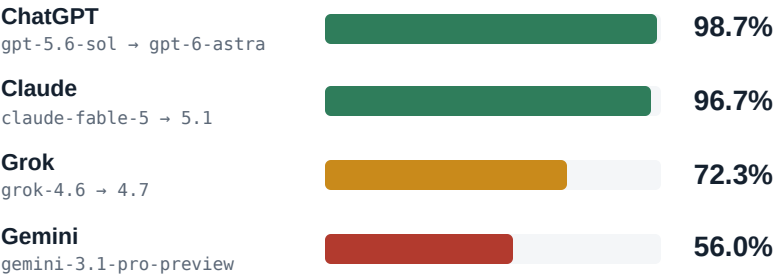


Even the top model fabricated a quote from the tax code.

The rest missed far more.

A closed-book benchmark of four frontier AI models on 25 CPA-level tax, GAAP and audit items, retested on every frontier release. Every response is graded by a licensed CPA against primary authority. Free to use in class.

CUMULATIVE ACCURACY · 150 RUNS PER MODEL



Each vendor's current flagship is tested every round; arrows show the version change. Grok: 102 correct of 141 answered, 9 runs timed out (68.0% if timeouts are scored as incorrect). Refusals are scored as incorrect.

WHAT STANDS OUT

2 vs. 66

Wrong answers out of 150: ChatGPT vs. Gemini, same questions

43%

Of Gemini's answers were wrong *and* marked high-confidence

Same 11

Gemini missed the identical 11 items, all three runs, in August and September

New ≠ better

Grok 4.7 repeated all five of 4.6's current-law misses; OpenAI's new flagship slipped from 100% to 97.3%

THE KEY FINDING

THE TOP MODEL'S ONLY MISSES · CHATGPT · ITEM C3

In 150 runs across two rounds, ChatGPT missed twice. Both misses came on the same question, and both were marked High confidence.

Asked which subparagraph of §642(b)(2) gives a complex trust its \$100 exemption, run 2 answered (B), quoting:
"Except as provided in subparagraph (C), all other trusts shall be allowed a deduction of \$100."

Actual (B): "A trust which, under its governing instrument, is required to distribute all of its income currently shall be allowed a deduction of \$300." The quote is 2002 wording plus a clause that never existed, presented as current law.

Run 1 gave the same wrong letter without a quote. Run 3 answered correctly: (A).

TRY IT IN CLASS: THE THREE PUBLIC ITEMS

A7 · §199A in 2026

Does §199A apply to tax years beginning in 2026, and what percentage is used in its general deduction formula?

Key: Yes; 20% (sunset repealed by P.L. 119-21)

A4 · SALT cap, 2026

For a taxpayer not married filing separately: the 2026 SALT limitation, the MAGI where it begins to phase down, and the floor.

Key: \$40,400; \$505,000 MAGI; \$10,000 floor

C3 · §642(b) (2)

Which subparagraph of IRC §642(b)(2) provides the \$100 personal exemption for a complex trust?

Key: Subparagraph (A). (B) is the \$300 rule for trusts required to distribute all income currently.

HOW THE TEST WORKS

25 items, 4 buckets Current law (10), disguised court holdings (5), known model traps (5), memos with a planted defect (5)	3 independent runs each Every run is its own API call: no thread, no prior context. Maximum reasoning, no tools, no voting
Pre-registered Question set and answer key SHA-256 committed before testing	Open logs Call log and full response text published as CSV

DISCUSSION PROMPTS

- Run the three items above on the model your students use. Do their results match ours? What would explain a difference?
- The same question, asked in three separate calls with no shared context, got two different answers. What does that imply for a workflow that asks once?
- The top model's only two misses in 150 runs came on the same question, both marked High confidence, and one quoted 2002 statute wording as current law. How would a reviewer catch an error the model gives no sign of?