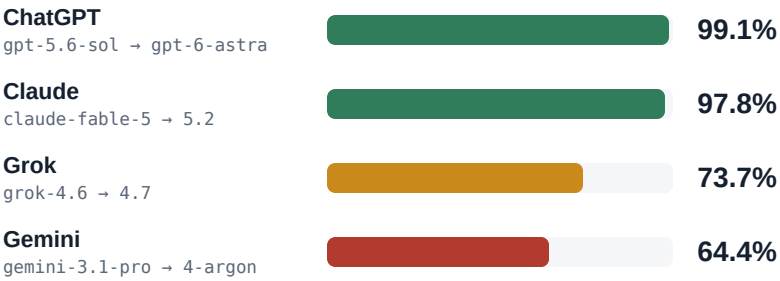
Two models got nothing wrong. One changed its answer.

Same 25 questions, three rounds, every frontier release.

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 · 225 RUNS PER MODEL



Arrows show the version change. Accuracy is correct answers over answers returned within 3 minutes: Grok 151 of 205 (20 runs unanswered in time), ChatGPT 220 of 222. Refusals score as incorrect.

WHAT STANDS OUT, NOVEMBER

0 wrong
Claude and ChatGPT, every answer they returned in November

56% → 81%
Google's new model, Gemini 4 Argon, versus its predecessor in September

3 of 10
Current-law items where Gemini was right on one run and wrong on another, same question, same settings

Unchanged
Grok 4.7 still answers five of the 2025 law changes as if the law never passed, as in August and September

THE KEY FINDING

SAME MODEL, SAME QUESTION, MINUTES APART · GEMINI 4 ARGON · ITEM A4

Asked for the 2026 SALT deduction limitation, run 1 answered, at low confidence:

"Under IRC § 164(b)(6) as enacted through my March 2025 knowledge cutoff, the $10,000 SALT deduction limitation ... expired after December 31, 2025 (leaving no SALT cap for 2026)."

Runs 2 and 3, at high confidence: **$40,400, phasing down above $505,000 MAGI, floor $10,000.** Correct.

It did the same on §199A (A7: "No" twice, then "Yes") and on the §179 limits. A user who asks once gets one of these answers and cannot tell which.

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Key: Yes; 20% (sunset repealed by P.L. 119-21)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

Committed in advance
Question set and answer key SHA-256 published in August, unchanged since

Open logs
Call log and full response text for the public items published as CSV

DISCUSSION PROMPTS

1. Run each of the three items above three times, in fresh chats, on the model your students use. Do the answers agree with each other? With the key?
2. One model knew a 2025 law change on one run and not the next. What does that imply for a workflow that asks once and moves on?
3. Two models now score at or near 100% on these 25 items. What would a harder test need to include, and how would you keep its answers out of training data?

Disclosure. ChatCPA sells a CPA exam test bank that AI drafts and a licensed CPA verifies. If models were useless the bank would be worthless; if they were flawless, CPA verification would be pointless. We profit only in the middle, so our incentives cut both ways.

Nicholas Miller, CPA · Oregon #14907
finance@chatcpa.io
chatcpa.io/ai-accuracy