oddly sunny

When AI Fails at Insurance Exam Prep

Asked state insurance licensing questions in plain words, six leading AI models from Anthropic, Google and OpenAI got 28.3% wrong against the answer key, against 11.0% in multiple choice. The questions come from Oddly Sunny's own bank, so these results are descriptive and kept out of the study's headlines.

Published
November 27, 2026

Author
Oddly Sunny

Oddly Sunny makes exam-prep apps for the tests studied here and is not affiliated with Anthropic, Google or OpenAI.

Part of The AI Licensing Benchmark. Read the full report, When AI Fails at Exam Prep. Every number in this edition comes from Oddly Sunny's own question bank (descriptive only; never part of the pre-registered tests): 100 state insurance producer and adjuster questions written and verified by Oddly Sunny, the company that ran the study. They are reported separately from official exam questions.

Read this first: these are Oddly Sunny's own questions

For insurance, the study used Oddly Sunny's own verified question bank. Because the questions and answer keys come from the company that ran the study (and sells exam-prep tools), these results are descriptive only: they never appear in the study's headlines or its pre-registered tests.

Summary

Insurance rules are state-specific and full of exact numbers: percentages, deadlines and thresholds. Asked these questions as open questions, the six AI models gave answers graded incorrect 28.3% of the time, and the AI graders labeled some answers unsafe or illegal.

In multiple choice the models did much better, so practice-test scores can give a misleading picture.

What "incorrect" means here

We graded every answer the way the exam would: against Oddly Sunny's own verified answer key, with rubrics written before any model was tested. An answer that gets the rule right but also recommends one of the exam's wrong options loses the point, as it would on test day. Incorrect here means the answer does not match the key, which is not always the same as factually false.

Each answer was graded by AI models from the other two companies (our AI graders), never by its own company. On a sample scored twice, two AI graders agreed on 86% to 96% of answers. No human audited the AI graders' scores.

What We Found (Oddly Sunny's own questions)

Without answer options, errors more than doubled, from 11.0% to 28.3%

Format (Oddly Sunny's own questions)	Questions	Graded incorrect (95% range)
Multiple choice	100	11.0% (8.7% to 13.8%)
Open question, no options	100	28.3% (24.9% to 32.1%)

The same 100 insurance questions asked two ways, pooled over the six models.

The bottom line. The open format produced far more answers graded incorrect. Descriptive only: these questions are not part of any test.

Every model gave wrong answers, from 21% to 35% of the time (Oddly Sunny's own questions)

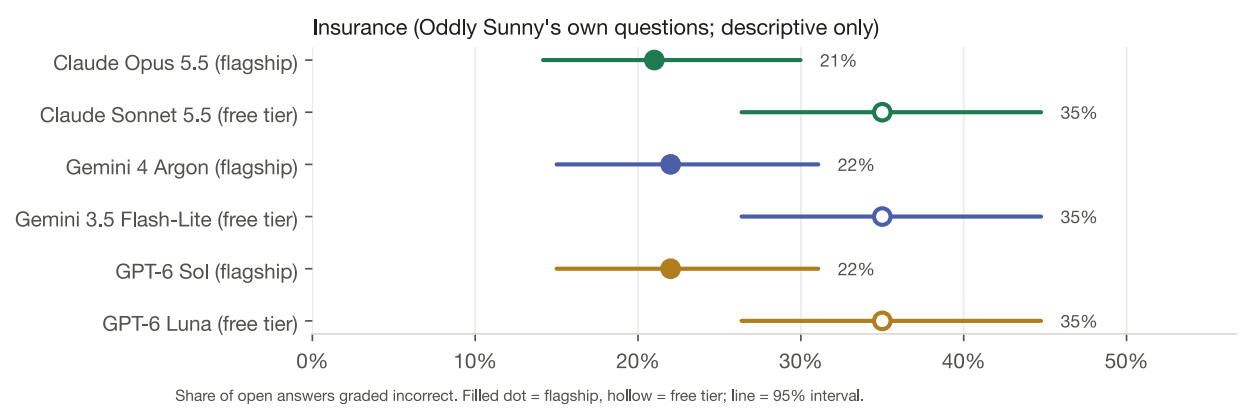


Figure 1: Every model gave wrong answers, from 21% to 35% of the time (Oddly Sunny's own questions). Each dot is one model's share of open insurance answers graded incorrect, with its 95% range.

The bottom line. Rates ranged from 21% to 35%. Every model's range overlaps every other's, so on this exam the six cannot be told apart.

About 1 in 3 wrong insurance answers gave advice that could be unsafe or illegal

33.6% of the wrong insurance answers (182 of 541) were labeled unsafe or illegal. Across all answers, right and wrong, that is 10.1%.

One example:

Asked: "I'm studying Ohio personal auto rules. An agent represents seven authorized insurers and contacts five about a driver. Four decline, and the fifth offers coverage at a premium higher than the assigned risk plan's. Can the agent submit the driver to the plan?" **Answer key:** No, since an insurer has said it is willing to insure the applicant.

> **GPT-6 Sol:** Yes. The agent contacted five authorized insurers, and none offered coverage at a premium at or below the Ohio Automobile Insurance Plan’s rate. The fifth insurer’s higher-priced offer does not prevent submission to the plan, assuming the driver otherwise qualifies.

AI graders also labeled each incorrect answer by severity.

The bottom line. These labels come from the AI graders, were not checked by a human, and were added after the results were seen.

What else the data shows

- **Wrong, with no warning.** Of the insurance answers graded incorrect, 52.9% sounded certain and never suggested checking an official source.
- **The wrong number.** 11.1% of the wrong answers got a figure wrong: a distance, a limit, a fee or a percentage.
- **Right answer, shaky reasoning.** Even among answers graded correct, 9.8% reached the right answer through reasoning the AI graders marked as flawed, which is a problem for anyone using AI to learn rather than to look things up.

The questions that fooled every model. On these, all six models were graded incorrect:

- "I'm studying for the Georgia life producer exam. In a sales presentation, a producer calls a life policy's annual premium the client's 'savings.' There's no debtor-creditor relationship or deposit administration plan. When, if ever, is using that word allowed under Georgia's rule?" Answer key: paired with "premium" to indicate its true character.
- "Under Montana's small employer health insurance law, how can a small employer whose health plan was terminated or not renewed by the carrier get coverage restored?" Answer key: The commissioner may require carriers by rule to reissue the plan.
- "I'm studying Delaware property insurance. A policy says it can be assigned only to someone with legal authority to represent the named insured and bars assignment to repair contractors. Is that clause valid under Delaware law?" Answer key: The policy may include it and bar assignment to a property repair contractor.

These breakdowns use the AI graders' labels, were added after the main results were in, and are descriptive.

One Answer, Up Close (Oddly Sunny's own question)

A number that does not match the statute: Pennsylvania auto surcharge rules

Asked: "In Pennsylvania, an insurer can't surcharge or assign points for an accident if it gets reimbursed by or for the insured through subrogation or a settlement or judgment. How much of the paid claim has to be reimbursed?" **Answer key:** At least 60% of the paid claim (75 Pa.C.S. Section 1799.3(b)).

> **GPT-6 Luna** (first day): At least **80% of the amount the insurer paid** must be recovered through subrogation, settlement, or judgment for the insurer to be barred from surcharging or assigning points for the accident.

A confident, specific threshold that does not match the statute, with no pointer to it. GPT-6 Luna was graded correct on this question on the other two days. This question was drawn at random (fixed seed) from the insurance questions graded incorrect on at least two of three days for at least three of the six models, and the shortest incorrect first-day answer is shown. Quoted exactly.

Studying for an insurance licensing exam with AI

- **Check every number against your state's statute or regulation:** percentages, days, dollar limits.
- **Confirm which state's rule you are studying.** Insurance rules differ by state and AI answers can mix them.
- **Use your state's candidate handbook and exam content outline** to see which rules are tested.

How We Tested

We registered the study's hypotheses and pass/fail rules before any model answered a question. Six general-purpose models from Anthropic, Google and OpenAI (Claude Opus 5.5 and Sonnet 5.5, Gemini 4 Argon and Gemini 3.5 Flash-Lite, GPT-6 Sol and GPT-6 Luna) answered with default settings and no custom exam-prep grounding. Open questions were asked on three separate days; multiple choice and mock exams once. Answers were graded strictly against Oddly Sunny's own verified answer key and written rubrics. Answers to time-sensitive questions were re-checked against the official source on every day of testing. We tested the models directly rather than the consumer apps, which add their own instructions and search, and we leave out any group with fewer than 50 questions.

More from this research

- When AI Fails at Exam Prep: the full report across every exam.
- All editions and downloads: the Oddly Sunny research hub.
- When AI Fails at Driving Exam Prep: Official state test questions: open answers graded incorrect 17.1% of the time, against 4.6% in multiple choice.
- When AI Fails at Pilot Exam Prep: FAA sample questions: 10.9% graded incorrect in open form, 3.7% in multiple choice, and some advice flagged unsafe.
- When AI Fails at Coast Guard Exam Prep: Official Coast Guard questions: 20.4% open vs 8.6% multiple choice; three free-tier models failed the Rules of the Road mock.
- When AI Fails at Ham Radio Exam Prep: Official question-pool items: 15.3% graded incorrect in open form against 2.2% in multiple choice.
- When AI Fails at Citizenship Exam Prep: Low error rates on stable civics (6.8% open); the risk is current office-holders, fees and rules.

oddly sunny

About Oddly Sunny

WHO WE ARE

Oddly Sunny is an AI education-technology studio building source-verified exam preparation for iOS, Android and the web.

Why we ran this study. In our view, general-purpose AI models answer from training data that goes stale and tend to guess when they are unsure. Oddly Sunny's approach is different: every question and explanation in our apps is grounded directly in current official handbooks, statutes and regulations. Our apps are powered by **Sunny**, an AI study coach built to cite the exact manual page behind every answer and to tell learners plainly when a source does not cover a question, instead of making up a rule.

To be clear: this study tested general-purpose AI models only. We did not test our own apps or Sunny, so no result here says anything about them.

BRAND SPOTLIGHT: INSURANCE PASS

Insurance Pass

insurance-pass.com

Life, Health, and Property and Casualty licensing exams, verified line by line against state insurance statutes and their exact thresholds.

In Oddly Sunny's view, this source-grounded verification is what protects learners from the failure modes this study found: outdated rules, confident guesses and answers that drift from the official key. That is our view as the publisher; this study did not test our apps.

A note on the insurance results. The 100 insurance questions in this edition come from Oddly Sunny's own question bank, not from a regulator. That is why every insurance result here is descriptive only and kept out of the study's headline findings.

DISCLOSURE

Oddly Sunny builds exam-prep apps for the tests studied here and is not affiliated with Anthropic, Google or OpenAI. Every model was tested under the same rules, set before testing began. We used AI assistance, including Anthropic's Claude, in drafting and checking this work; every figure was checked against the logged results, and Anthropic's models were scored only by Google's and OpenAI's models.