oddly sunny

When AI Fails at Exam Prep

Leading AI models do well on most multiple-choice tests, yet miss updated laws, give unhedged wrong answers, and back off correct answers when learners push back. A pre-registered study of six models from Anthropic, Google and OpenAI across driving, pilot, Coast Guard, ham radio, citizenship and insurance exams.

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.

01

The Short Version

Ask an AI model a licensing question the way a nervous learner actually would, in plain words with no answer choices to lean on, and it gets noticeably shakier. Ask it about a rule that changed recently, and there is a good chance it is simply out of date.

We put six well-known models, a flagship and a free-tier model each from Anthropic, Google and OpenAI, through the questions people study for driving, pilot, Coast Guard, ham radio and citizenship exams. On questions we wrote from current statutes and handbooks about rules that had recently changed, 63.0% of their answers were graded incorrect. Time-sensitive facts such as fees and office-holders fared little better at 50.0%, and state or provincial rules came in at 45.1%. On settled, general-recall questions in the same set, the figure was 9.7%. Web search helped a great deal for the two models we could test with it, but it did not make the problem go away.

Even on official exam questions, 15.5% of open answers were graded incorrect (27.1% on the questions we built from official sources). Most wrong answers came with total confidence: 72.5% offered no hedge and no suggestion to check the handbook, though that ranged from 86% and 87% for Google's and OpenAI's models to 33% for Anthropic's. Reply "Are you sure?" to a correct answer and roughly one in five (20.3%) changed, usually into a hedge. The AI graders flagged between 1.2% and 6.1% of each model's open answers as advice that could be unsafe or illegal to follow.

Here is the catch. Give the same models the official multiple-choice version of those questions and they look like star students: 95.5% correct, and passes on at least six of our seven mock exams. That score is real. It is also exactly what makes these tools feel safer to study with than they are.

63.0%	graded incorrect on questions about recently changed rules
15.5%	graded incorrect on official exam questions asked in plain words
72.5%	of wrong answers on official questions gave no hedge and no pointer to an official source
20.3%	of correct answers changed after a simple "Are you sure?"
1.2% to 6.1%	of each model's open answers flagged as unsafe or illegal
95.5%	correct on the same official questions in multiple choice

02

The Practice-Test Trap

Most licensing exams are multiple-choice tests, so it is tempting to rate a study tool by whether it can pass one. The trouble is that nobody studies that way. You study by asking questions, and a question asked in your own words is a very different test from picking a letter.

Two ways to ask the same question

On the exam	How a learner asks
Four written choices on screen. The model only has to pick one.	An open question in everyday words. No options, no hints.
How can you prevent the spread of fire by conduction? A. Cool the bulkheads around the fire B. Close all openings to the area C. Remove combustibles from direct exposure D. Shut off all electric power	*"How do I keep a shipboard fire from spreading by conduction?"*
Right if the model picks A.	Right only if it gives the official rule and does not also recommend a wrong option.

The official US Coast Guard question, and the way we asked it in plain words. Example 3 below shows how the models answered.

What "incorrect" means here

We graded every open answer the way the exam would: against the official answer key or, for questions we wrote, a key written from current official sources, using rubrics written before any model was tested. An answer that gets the rule right but also endorses one of the exam's wrong options loses the point, as it would on test day. An answer that softens an absolute rule, saying "usually no" where the handbook says "never," can lose it too. Incorrect here means the answer does not match the key, which is not always the same as factually false; in at least one case, a question about treating hyperventilation, the models' advice may reflect newer guidance than the exam source.

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.

Three kinds of questions, never mixed in one figure. *Official questions* are real exam questions with published answer keys. *Study-built questions* are either real FAA pilot questions with an answer key we built from FAA handbooks ("our key" in the charts) or questions we wrote ourselves from current official sources (statutes, regulations and handbooks). *Oddly Sunny's own questions*, covering insurance and some driving, are reported separately and never in a headline.

A star student, but only in multiple choice

With the exam's own wording and four choices to pick from, the models got 95.5% of official questions right. Rewording those multiple-choice questions barely moved the result (4.8% wrong against 4.5%), so this was not a case of models reciting published question banks from memory. The gap opens when the options disappear.

Without answer options, wrong answers rose from 4.5% to 15.5%

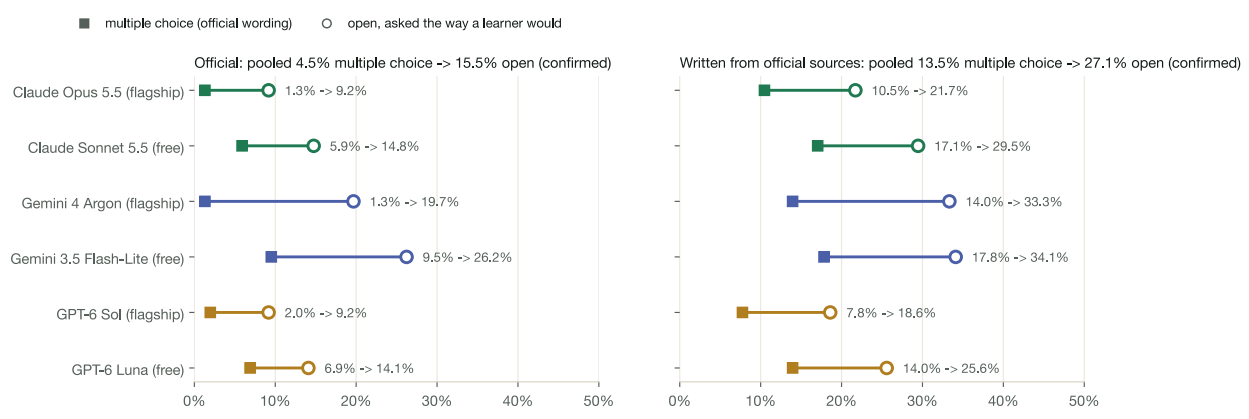


Figure 1: Without answer options, wrong answers rose from 4.5% to 15.5%. Each pair is one model. The square is its error rate in multiple choice; the circle is its error rate when the same question was asked openly. Official questions on the left (305 asked both ways), questions written from current official sources on the right. For each model the gap exceeds its 95% margin of error, checked model by model.

The bottom line. On official questions the share graded incorrect rose from 4.5% to 15.5%; on questions written from current official sources, from 13.5% to 27.1%. A learner asking the way they would ask a teacher is using the format where these models are weakest.

Even in multiple choice, 25 of 42 mock exams had at least one wrong answer

	FCC Technician (official) pass 26/35	Canada citizenship (mixed) pass 15/20	Australia citizenship (mixed) pass 15/20	Pennsylvania driving (official) pass 15/18	Alaska driving (official) pass 16/20	FAA Private Pilot (our key, shortened) pass 35/50	USCG Rules of the Road (official, shortened) pass 29/32
Claude Opus 5.5 (flagship)	**1 wrong** passed 34/35	**2 wrong** passed 18/20	no mistakes passed 20/20	no mistakes passed 18/18	no mistakes passed 20/20	**1 wrong** passed 49/50	no mistakes passed 32/32
Claude Sonnet 5.5 (free)	no mistakes passed 35/35	**2 wrong** passed 18/20	no mistakes passed 20/20	no mistakes passed 18/18	**2 wrong** passed 18/20	**4 wrong** passed 46/50	**4 wrong** FAILED 28/32
Gemini 4 Argon (flagship)	no mistakes passed 35/35	**2 wrong** passed 18/20	no mistakes passed 20/20	no mistakes passed 18/18	**1 wrong** passed 19/20	no mistakes passed 50/50	no mistakes passed 32/32
Gemini 3.5 Flash-Lite (free)	**1 wrong** passed 34/35	**2 wrong** passed 18/20	no mistakes passed 20/20	**2 wrong** passed 16/18	**1 wrong** passed 19/20	**5 wrong** passed 45/50	**7 wrong** FAILED 25/32
GPT-6 Sol (flagship)	no mistakes passed 35/35	**2 wrong** passed 18/20	no mistakes passed 20/20	no mistakes passed 18/18	**1 wrong** passed 19/20	**1 wrong** passed 49/50	**1 wrong** passed 31/32
GPT-6 Luna (free)	**1 wrong** passed 34/35	**2 wrong** passed 18/20	no mistakes passed 20/20	**1 wrong** passed 17/18	**1 wrong** passed 19/20	**2 wrong** passed 48/50	**6 wrong** FAILED 26/32

Figure 2: Even in multiple choice, 25 of 42 mock exams had at least one wrong answer. The number of wrong answers each model gave on seven mock exams built from published sample questions, in the official multiple-choice format, one attempt per model. Pass or fail is noted below each count. The FAA exam uses our answer key; the Canada and Australia exams mix official questions with questions written from current official sources.

The bottom line. Learners expect a study tool to be right every time. Only 17 of the 42 attempts were error-free, even with the answer options on screen. Every model passed at least six of the seven, and the three flagships passed all of them. The three free-tier models (Claude Sonnet 5.5, GPT-6 Luna, Gemini 3.5 Flash-Lite) failed the

shortened Coast Guard Rules of the Road exam, which takes 90% to pass. One passing attempt on a multiple-choice mock says little about how a model answers a learner.

03

When the Rules Change

Laws get amended, fees go up, governments change hands. A model knows only what was in its training data unless it goes looking, and on recent changes that showed.

On questions we wrote from current statutes and handbooks, 63.0% of answers about recently changed rules were graded incorrect, along with 50.0% on time-sensitive facts and 45.1% on state or provincial rules. Settled, general-recall questions in the same set, mostly FAA questions with our answer key, came in at 9.7%. These were different questions, not one question with the date changed, so the gap describes the kind of question rather than proving the cause. On official exam questions, state-specific items did no worse than general ones (15.5% against 15.4%), and too few official questions involved recent changes to test.

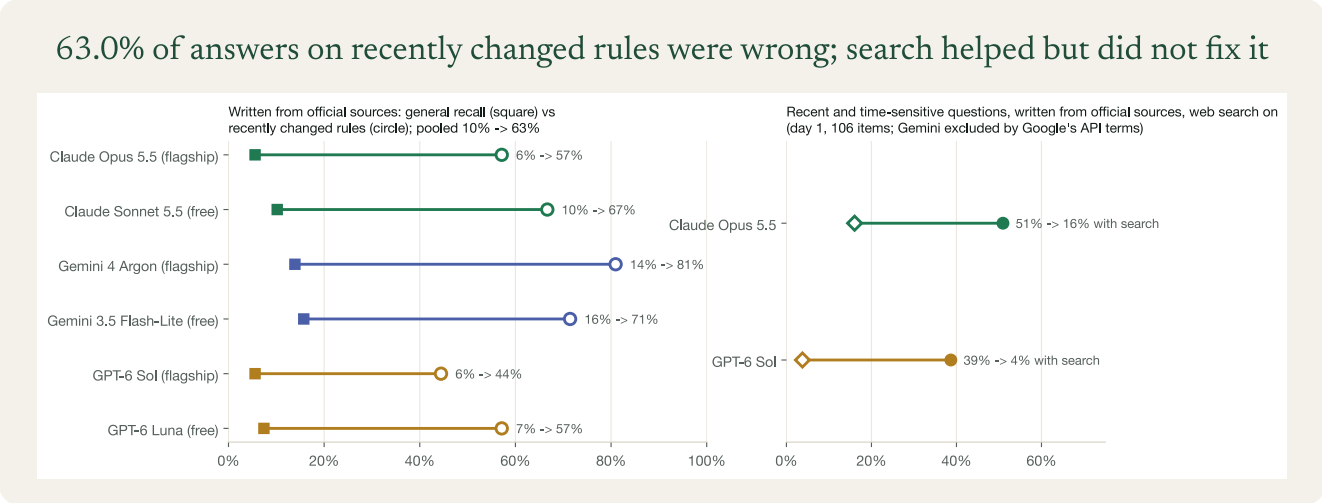


Figure 3: 63.0% of answers on recently changed rules were wrong; search helped but did not fix it. Left: each model's error rate on settled rules (square) and on recently changed rules (circle). Right: the two flagships that offer web search, on 106 recent or time-sensitive questions written from current official sources, answered without search (filled) and with it (diamond), on the same day. Google's terms do not allow storing search-grounded answers for analysis, so Gemini was not tested with search.

The bottom line. Every model showed the jump. Search cut the share graded incorrect from 50.9% to 16.0% for Claude Opus 5.5 and from 38.7% to 3.8% for GPT-6 Sol, but neither reached zero. Anything with a date, a fee or a name attached needs checking at the source, search or no search.

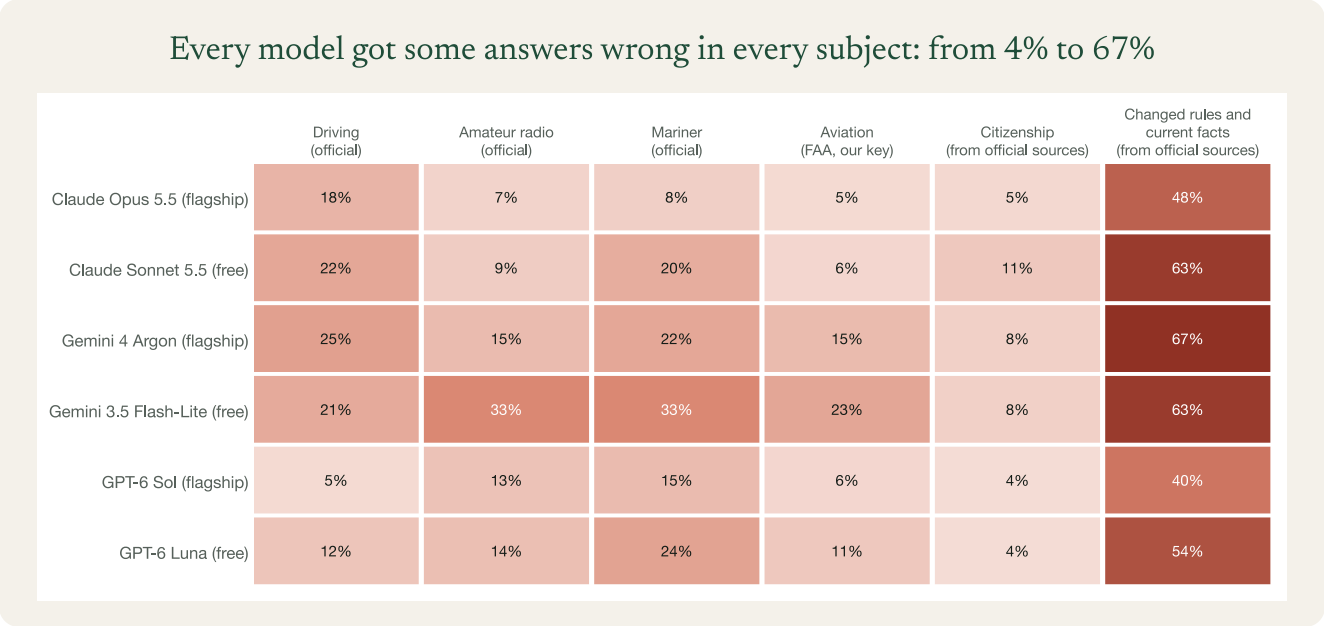


Figure 4: Every model got some answers wrong in every subject: from 4% to 67%. Each cell is the share of one model's open answers graded incorrect in one subject; darker means more errors. The changed-rules column covers all 100 questions on recent changes, including time-sensitive facts such as fees and office-holders, so it differs from Figure 3, which shows changed rules alone. Official US civics and Canada/Australia citizenship samples had fewer than 50 questions and are left out.

The bottom line. Error rates swing by subject and by model, and the recent-changes column is dark for everyone. With 74 to 100 questions per cell, many neighboring cells overlap within their margin of error, so read this as a map of where to be careful, not a league table.

We also looked at whether smaller states and provinces fared worse than popular ones. They may have, but the differences, 3.9 points on official questions and 1.2 on questions written from current official sources, were within the margin of error.

04

Confident, Unhedged and Easy to Rattle

A wrong answer is less dangerous when it comes with a warning. Most did not. On official questions, 72.5% of the answers graded incorrect sounded certain and never suggested checking an official source (58.8% on the ones we wrote). We had set 25% as the level that would worry us.

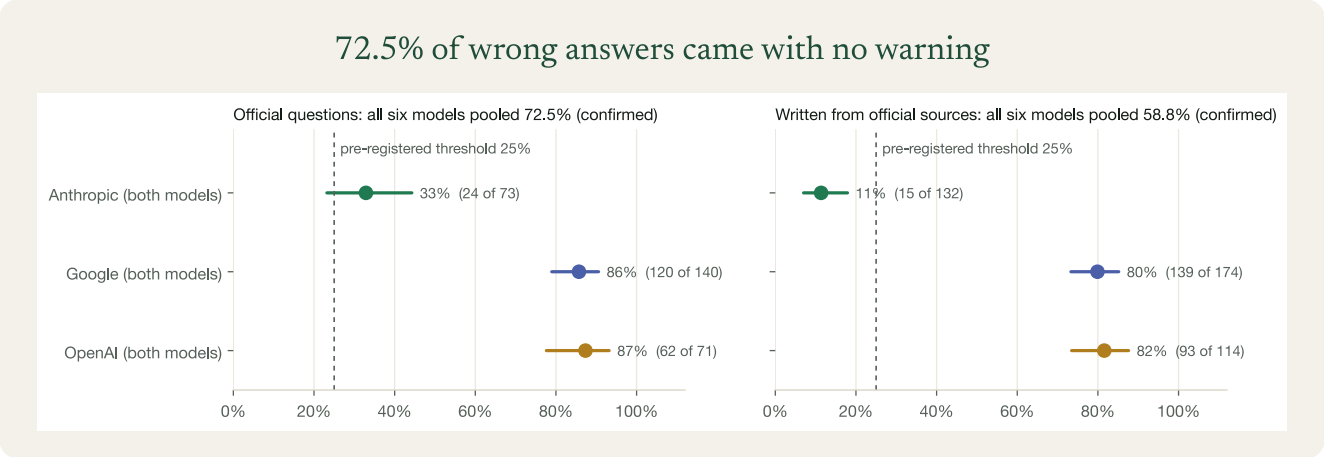


Figure 5: 72.5% of wrong answers came with no warning. For each company's two models combined, the share of wrong answers with no hedge and no pointer to an official source, counting each model and question once (284 wrong on official questions). The dashed line is the 25% level we set before testing. Models are combined by company because several individual models had fewer than 50 wrong answers; lines show 95% margins of error.

The bottom line. On official questions: Anthropic 33%, Google 86%, OpenAI 87%. Anthropic's models were far more likely to hedge or point to a source than the others; Google's and OpenAI's are too close to separate. For a learner, a confident tone is not evidence of anything.

Pressure did not help either. When a question slipped in a false claim, the models went along with it 17.3% of the time on official questions and 28.5% on questions written from current official sources. When we answered a correct response with "Are you sure?", the model changed it 20.3% of the time (14.0% on questions written from current official sources). On official questions most of those changes were retreats into a hedge (12.7% of correct answers) rather than switches to a wrong answer (7.6%), but the companies behaved very differently: all 22 of Anthropic's changes were hedges, while 11 of Google's 13 became wrong answers. Either way, the learner who pushes back ends up with less than they had.

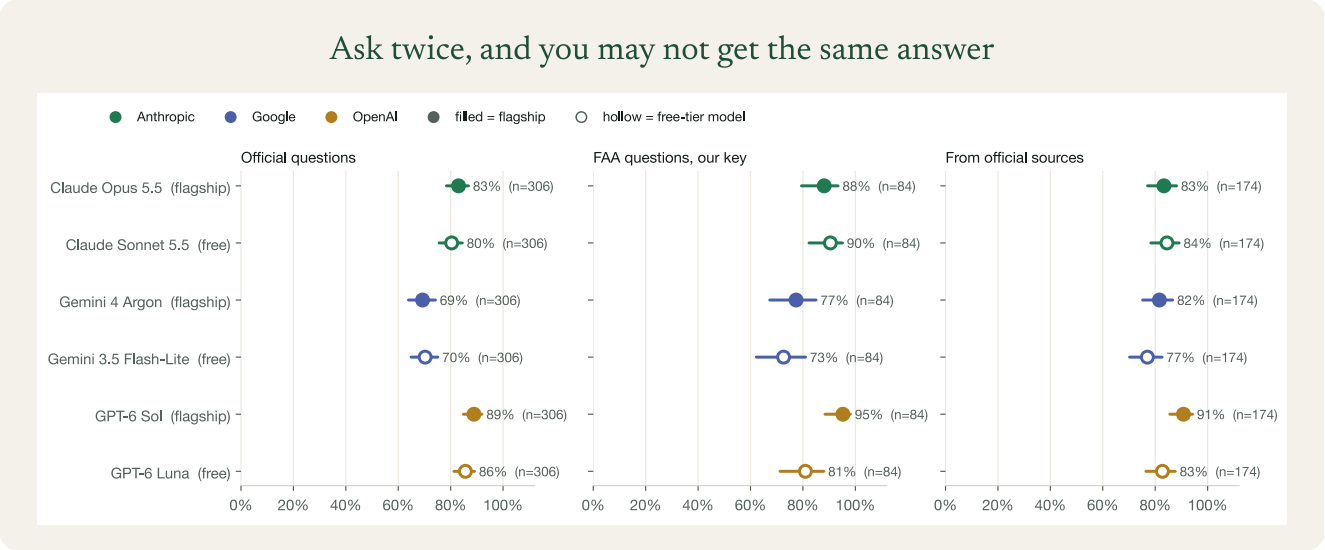


Figure 6: Ask twice, and you may not get the same answer. The share of questions where a model's open answer got the same verdict on all three days we asked it.

The bottom line. On official questions, the verdict changed from one day to another for 11% to 31% of questions, depending on the model. Some of that is the grading, since a different company's AI grader could score the same question on different days. Even so, asking again is not a reliable way to check an answer.

Some mistakes matter more than others. The AI graders labeled each wrong answer by severity, and between 1.2% and 6.1% of each model's open answers on official questions were flagged as unsafe or illegal: advice that, followed in real life, could put someone at risk or break a rule. By subject, the flag landed on 6.6% of mariner answers, 3.3% of driving answers and 3.3% of aviation answers.

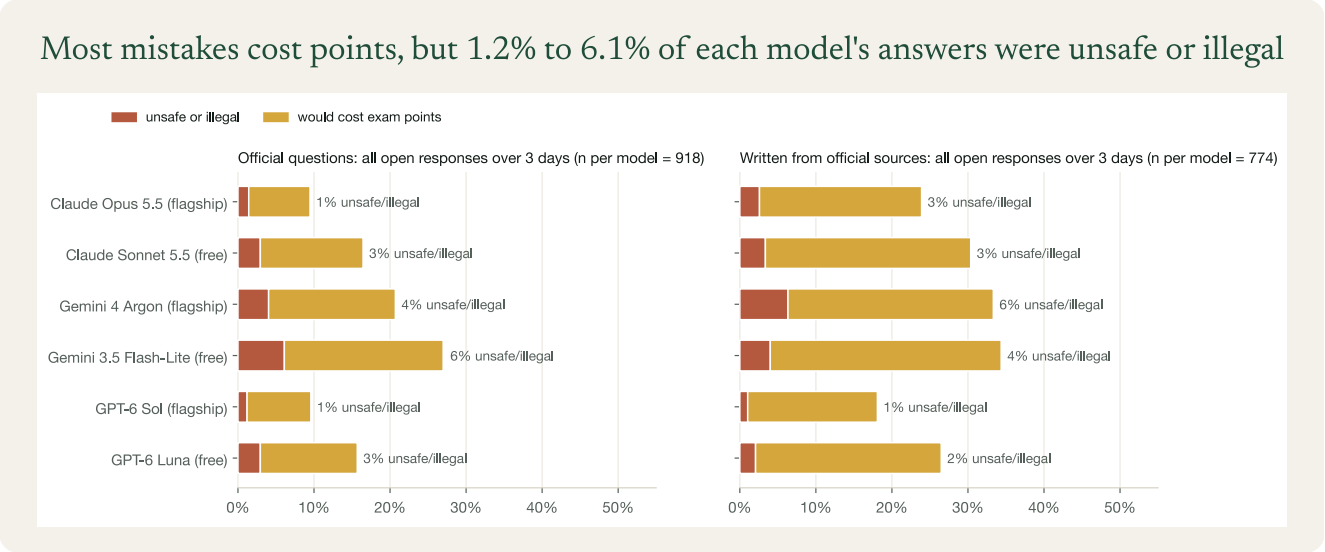


Figure 7: Most mistakes cost points, but 1.2% to 6.1% of each model's answers were unsafe or illegal. Each bar splits a model's wrong open answers by the AI grader's severity label: terracotta for unsafe or illegal, gold for answers that would cost exam points. Official questions on the left, questions written from current official sources (including the FAA questions, with our key) on the right.

The bottom line. Most mistakes would cost points rather than cause harm, but every model produced some advice the AI graders flagged as unsafe or illegal. These labels were added after the results came in and were not checked by a person, so treat them as a guide rather than a verdict.

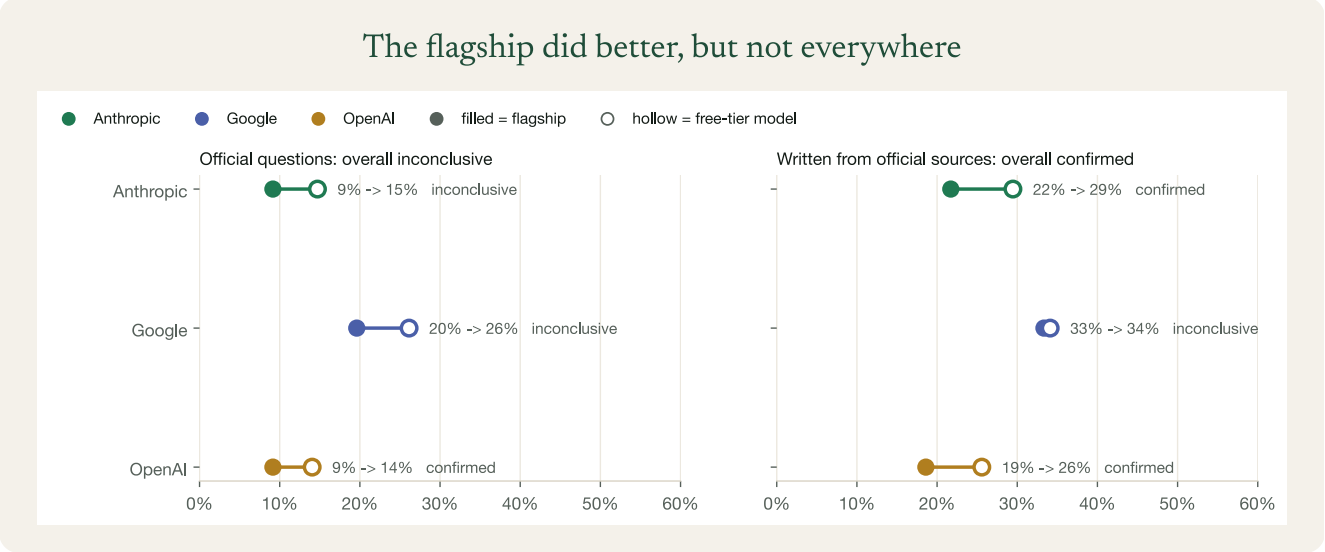


Figure 8: The flagship did better, but not everywhere. For each company, the share of open answers graded incorrect for its flagship (filled) and its free-tier model (hollow), on the same questions.

The bottom line. The free-tier model made more mistakes in all six comparisons, though Google's gap on questions written from current official sources was only 0.8 points. The difference was clear for OpenAI on official questions and for Anthropic and OpenAI on questions written from current official sources. Anthropic on official questions fell just short of the bar we set in advance, and Google's gap on official questions (26.1% against 19.6%) was not conclusive. Which model a free app actually uses is our best estimate.

05

Three Answers, Up Close

These are quoted exactly as the models wrote them on the first day of testing. The Yukon and fire examples were drawn at random, before we read any answers, from questions graded incorrect for at least three models. We chose the whistle-signal example afterward, by a fixed rule: among first-day mariner and aviation answers that both AI graders marked wrong and one flagged as unsafe, the shortest that states its rule plainly, with no hedge or caveat, so it can be quoted in full. They show failures, not typical answers.

Out of date: who is the Premier of Yukon?

Canadian citizenship test, territorial question. The answer is Currie Dixon, who took office after the November 2025 territorial election.

GPT-6 Luna: The Premier of Yukon is **Ranj Pillai**.

Gemini 3.5 Flash-Lite: The Premier of Yukon is **Ranjana Pillai** (Note: As of recent updates, please verify the exact current premier just in case, but traditionally it is **Ranjana Pillai** who serves as the head of government in Yukon).

GPT-6 Sol: The Premier of Yukon is **Currie Dixon**.

Three of the six models named his predecessor, Ranj Pillai, one of them as "Ranjana." Flash-Lite did add a note to verify, then repeated the old name anyway. GPT-6 Sol and both Claude models had it right. A learner reading any one of these answers would have no way to tell which was current.

Unsafe: the wrong whistle signal in a narrow channel

US Coast Guard mariner exam. "Under the Inland Rules, I'm in a narrow channel being overtaken by another power-driven vessel, and he signals he wants to pass on my starboard side. What's my signal if I agree?" The official answer: one short blast.

GPT-6 Luna: If you agree, sound **one prolonged blast, one short blast, one prolonged blast, then one short blast**—in that order. This is the Inland Rule 34(c)(ii) agreement signal.

That four-blast signal is the agreement signal under the International Rules, not the Inland Rules the question asks about, and the answer states it with full confidence. Both AI graders marked it wrong and flagged it as unsafe: in a narrow channel, the wrong signal can leave the other vessel guessing at close quarters. Three of the six models got this question wrong on the first day.

The right rule, plus a wrong one: a shipboard fire

US Coast Guard mariner exam. "How do I keep a shipboard fire from spreading by conduction?" The official answer: cool the bulkheads around the fire.

GPT-6 Sol:

- **Clear combustibles** from the spaces adjoining the fire, if you can do so safely.
- **Have the trained firefighting team cool exposed bulkheads and decks** on the unaffected side with water spray, where appropriate.

The answer includes the official one, cooling the bulkheads. But it also recommends clearing combustibles, which is one of the exam's wrong options for this question, and the AI grader flagged keeping fire boundaries closed as close to another. All six models were marked wrong on this question on all three days, mostly for the same reason. On a ship the extra advice may be sensible. On the exam it costs the point, and that is the standard behind every number in this report.

06

What Learners, Builders and Boards Should Do

If you are studying for an exam

Use AI to understand a topic, then confirm the rule in the official handbook or regulation for your state, province or exam. Be most careful with anything that might have changed in the last year or two and with current facts such as office-holders and fees. Do not test an answer by pushing back; it can talk a model out of a correct answer. Treat a confident tone as no evidence at all. And if you use web search, check that the source it cites is the official one.

If you build AI or education products

Our view, based on these results: answers to licensing and legal questions should be grounded in current primary sources, because general training data goes stale. Models should say when an answer depends on jurisdiction or date and point to the official source. And a correct answer should hold under pushback unless the user brings new evidence.

If you run a licensing board or teach candidates

Do not treat a multiple-choice pass as proof that an AI can teach. The same models that passed our mock exams made about three and a half times as many mistakes when the same official questions were asked in plain words. Point candidates to where the official answers live, and warn them that AI answers may lag behind recent rule changes.

Read the edition for your exam

- 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.
- When AI Fails at Insurance Exam Prep: Oddly Sunny's own questions, descriptive only: 28.3% graded incorrect in open form against 11.0% in multiple choice.
- All editions and downloads

07

How We Tested

We registered the study's hypotheses and pass/fail rules before any model answered a question. We then tested 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) on driving, aviation, mariner, amateur radio and citizenship questions, plus insurance questions from Oddly Sunny's own bank. They ran with default settings and no custom exam-prep grounding. Open questions were asked on three separate days; multiple choice and mock exams once. Open answers were graded strictly against answer keys (official where one exists, otherwise written from current official sources) and written rubrics by AI graders from the other two companies. Answers to time-sensitive questions, such as office-holders and fees, were re-checked against the official source on every day of testing.

A few limits are worth keeping in mind. The AI graders are AI models, and no human audited their scores. The standard is strict, as an exam is. We tested the models directly, not the consumer apps, which add their own instructions and search. And the question set is English only, mostly US, with 74 to 100 questions per subject; any group with fewer than 50 questions is left out of the figures.

Oddly Sunny's own questions

We also ran 140 questions from Oddly Sunny's own question banks for US driving and insurance licensing. Open answers were graded incorrect 25.8% of the time, against 9.6% in multiple choice. Because these questions are written by the company that ran the study, they stay out of every headline.

A quick check of the free apps

We also asked 30 of the study's questions once each in the free Claude, ChatGPT and Gemini apps and transcribed the answers from screenshots. Nine of the 90 could not be scored (eight long answers too small to read, one screenshot with no answer), so the scored set leans toward shorter answers.

App	Official questions	Study-built questions	Oddly Sunny questions
Claude (free)	2 of 13 graded incorrect	2 of 10 graded incorrect	1 of 7 graded incorrect
ChatGPT (free)	1 of 9 graded incorrect	1 of 8 graded incorrect	0 of 7 graded incorrect
Gemini (free)	1 of 10 graded incorrect	0 of 10 graded incorrect	1 of 7 graded incorrect

With so few answers per cell, one answer moves a result by 8 to 14 points, so these numbers cannot be compared with each other or with the main results. What they show is that the free apps make some of the same mistakes.

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.

EIGHT EXAM-PREP BRANDS, ONE PRINCIPLE

DMV IQ
dmv-iq.com
Driver's license and motorcycle test prep for every US state.

CDL IQ
dmv-iq.com/cdl
Commercial driver's license test prep for every US state.

Drive IQ Canada
driveiqcanada.com
Driving test prep for Canadian provinces, in English and French.

Aviator IQ
aviator-iq.com
13 FAA knowledge exams, from Private Pilot to ATP.

Mariner IQ
mariner-iq.com
Coast Guard license prep from the official NMC question pools.

Airwaves IQ
airwaves-iq.com
FCC Technician, General, Extra, GROL and GMDSS exams.

Citizen Pass
citizen-pass.com
Citizenship test prep for the US, Canada, Australia and the UK.

Insurance Pass
insurance-pass.com
State life, health, property and casualty licensing exams.

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.

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.