oddly sunny

When AI Fails at Driving Exam Prep

Six leading AI models from Anthropic, Google and OpenAI pass state driving mock exams. Ask the way new drivers ask and they get nearly four times as many wrong against the official answer key, sometimes soften rules the handbook states as absolutes, and miss more than half of the questions on recently changed rules.

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. This edition covers 85 official knowledge-test questions from 11 US states (Alaska, California, Delaware, Iowa, Maryland CDL, Michigan, Montana, Nevada, New York, Pennsylvania, Utah) with the official answer key.

Summary

Driving rules differ by state, and handbooks state many rules in absolutes: never, always, within a set distance. When we asked official state test questions as open questions, the six AI models gave more answers graded incorrect, and some answers were labeled unsafe or illegal. The example below shows one way it happens: softening a rule the handbook states as absolute.

Multiple-choice scores and passing mock exams look reassuring, and that is the trap: they measure a different, easier task.

What "incorrect" means here

We graded every answer the way the exam would: against the official 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

Without answer options, errors rose from 4.6% to 17.1%

Format	Questions	Graded incorrect
Official multiple choice	84	4.6% (3.1% to 6.8%)
Open question, no options	85	17.1% (14.0% to 20.6%)

The same driving questions asked two ways, pooled across the six models: as the official multiple-choice item, and as an open question with no options. Ranges are 95% margins of error.

The bottom line. Take the options away and the error rate climbs from 4.6% to 17.1%. Across all subjects this gap held up in the study's main test; here it is shown for this exam alone.

Every model gave wrong answers, from 5% to 25% of the time

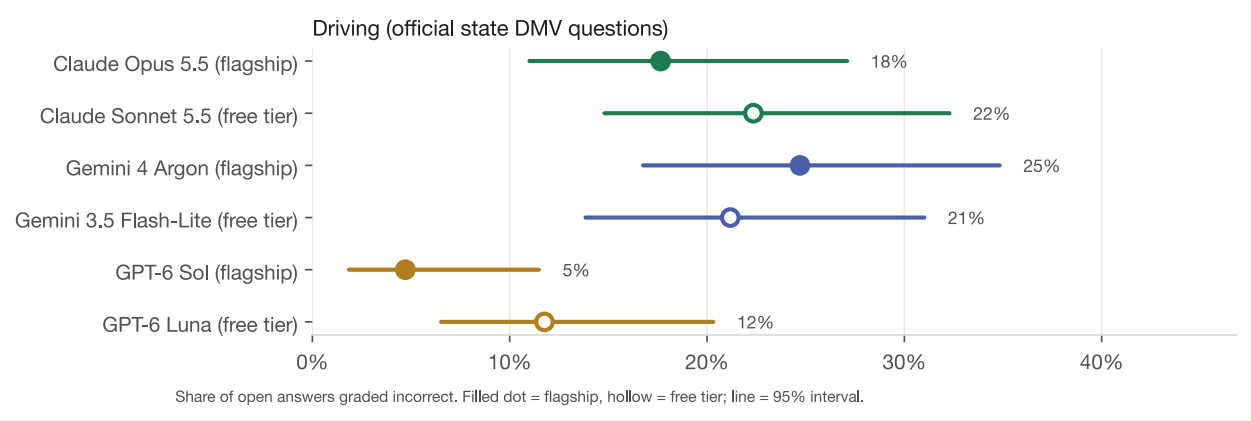


Figure 1: Every model gave wrong answers, from 5% to 25% of the time. Each dot is one model's share of open driving answers graded incorrect; the line is its 95% margin of error.

The bottom line. Rates ran from 5% to 25%. The only clear differences: Claude Sonnet 5.5 against GPT-6 Sol; Gemini 4 Argon against GPT-6 Sol; Gemini 3.5 Flash-Lite against GPT-6 Sol. The rest overlap, so they are not a ranking.

About 1 in 5 wrong answers gave advice that could be unsafe or illegal

When a model got a driving question wrong, the mistake was not always just a lost exam point. **18.5%** of the wrong answers (51 of 276) were flagged as unsafe or illegal: advice that, followed in real life, could put someone at risk or break a rule. Across all answers, right and wrong, that is 3.3%.

The AI graders also labeled each wrong answer by how much harm it could do.

The bottom line. These flags come from the AI graders, were added after the results came in and were not checked by a person, so treat them as a guide rather than a verdict.

Even in multiple choice, 7 of 12 attempts on the Pennsylvania and Alaska mock exams had at least one wrong answer

Exam (pass mark)	Opus 5.5	Sonnet 5.5	Argon	Flash-Lite	GPT-6 Sol	GPT-6 Luna
Pennsylvania driving, 18 questions (15 of 18)	no mistakes 18/18, passed	no mistakes 18/18, passed	no mistakes 18/18, passed	**2 wrong** 16/18, passed	no mistakes 18/18, passed	**1 wrong** 17/18, passed
Alaska driving, 20 questions (16 of 20)	no mistakes 20/20, passed	**2 wrong** 18/20, passed	**1 wrong** 19/20, passed	**1 wrong** 19/20, passed	**1 wrong** 19/20, passed	**1 wrong** 19/20, passed

Opus 5.5, Argon and GPT-6 Sol are the flagships; Sonnet 5.5, Flash-Lite and GPT-6 Luna the free-tier models.

Full-length state knowledge tests built from official sample questions, in multiple-choice format, one attempt each.

The bottom line. Every model passed both. A pass here is a multiple-choice result and says little about how a model handles a new driver's open question.

Keeping up with change

State driving rules change all the time: fines, points, crash-report thresholds and penalties. AI models learn from data collected up to a cutoff date, and nothing tells them when a state rewrites its handbook.

We wrote 56 questions from state driving laws and DMV rules that changed recently. On those, **55.4%** of answers were graded incorrect, far above the rate on the official test questions above. If your state changed a rule in the last year or two, assume a chatbot may not know.

A few of the questions we asked:

- "In Florida, if I'm in a crash where nobody is hurt, how much damage does there have to be before I must report it to police right away?" **Current answer:** $2,000 (changed October 2026). **6 of 6 models got it wrong.**
- "In Nevada, what's the minimum jail time for a second DUI within 7 years?" **Current answer:** 20 days (changed October 2025). **6 of 6 models got it wrong.**
- "In New York, how many points go on your driving record for passing a stopped school bus?" **Current answer:** 8 (changed February 2026). **6 of 6 models got it wrong.**

What else the data shows

- **Wrong, with no warning.** Of the driving answers graded incorrect, 62.0% sounded certain and never suggested checking an official source.
- **The wrong number.** 15.9% 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, 1.8% 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:

- "In California, is parking in a crosshatched (diagonal lines) area ever allowed?" Answer key: It is never allowed to park in a crosshatched (diagonal lines) area.

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

Reading the figures

Some driving questions came with an official picture, such as a road sign or a hand signal, and we gave the models the same image. Every model got those right on most of the days we asked.

One Answer, Up Close

A hedged absolute: crosshatched parking areas in California

Asked: "In California, is parking in a crosshatched (diagonal lines) area ever allowed?" **Official answer:** it is never allowed to park in a crosshatched area.

> **GPT-6 Sol** (first day): **Usually no.** If the crosshatched area is the access aisle next to a disabled parking space, California law prohibits parking or leaving a vehicle there **even if you have a disabled placard or plate** (Vehicle Code § 22507.8).
>
> For crosshatched areas elsewhere, the lines alone don’t establish a statewide rule for every location. Follow the posted signs and applicable parking rules—but don’t treat a striped area as a parking space unless it’s designated for parking.

The answer gets the disabled-aisle rule right but opens with "Usually no" where the handbook says "never", so it was scored incorrect. All six models were graded incorrect on this question on at least two of three days, mostly for hedging or adding exceptions. Quoted exactly.

Studying for your knowledge test with AI

- **Use your own state's handbook as the final word.** Rules differ by state, and the test follows the handbook's wording.
- **Watch for softened absolutes.** If the handbook says "never" or "always", an answer that says "usually" can cost you the point.
- **Check numbers yourself:** distances, speeds, fines and deadlines are where small errors hide.

- **Do not treat a practice-test pass from an AI as proof it can teach you.**

A separate note: Oddly Sunny's own driving questions

The study also tested 40 driving questions from Oddly Sunny's own question bank (Texas, Florida, Ohio and Georgia). That is fewer than 50 questions, so no number is reported, and because the bank belongs to the company that ran the study, these questions are never part of any headline or test.

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 official answer keys 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 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.

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: DMV IQ AND DRIVE IQ CANADA

DMV IQ

dmv-iq.com

Driver's license, motorcycle and CDL test prep for all 50 US states and DC, mapped to each state's current driver manual.

Drive IQ Canada

driveiqcanada.com

Driving test prep for Canadian provinces, in English and French.

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.