oddly sunny

# When AI Fails at Pilot Exam Prep

Six leading AI models from Anthropic, Google and OpenAI pass a shortened Private Pilot mock exam. Ask FAA knowledge-test questions in a pilot's own words and they get about three times as many wrong against an answer key built from FAA sources, and some of the advice is unsafe.

**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 FAA sample knowledge-test questions (Private, Commercial, Instrument and small unmanned aircraft), 84 of them also asked as open questions. The FAA does not publish answers for these samples, so the answer key was built by the study from FAA handbooks and regulations; these questions are reported separately from questions with an official key.*

# Summary

Pilots study procedures where the exact wording of the handbook matters. When we asked FAA sample questions as open questions, the six AI models gave more answers graded incorrect than in multiple choice, and the AI graders labeled some answers unsafe or illegal, including advice that contradicts the FAA's own guidance on contaminated runways. Rules that changed recently (for example medical and registration rules) were part of our separate questions on recently changed rules, written from current government sources, where errors were highest overall, but there were too few aviation questions in it to report on their own.

Every model passed our shortened Private Pilot mock exam, which is the false sense of security this report is about.

### What "incorrect" means here

We graded every answer the way the exam would: against an answer key the study built from FAA handbooks and regulations, 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 3.7% to 10.9%

| Format | Questions | Graded incorrect |
|---|---|---|
| Multiple choice (FAA wording) | 85 | 3.7% (2.4% to 5.8%) |
| Open question, no options | 84 | 10.9% (8.5% to 13.9%) |

The same aviation questions asked two ways, pooled across the six models: as the FAA 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 3.7% to 10.9%. 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 23% of the time

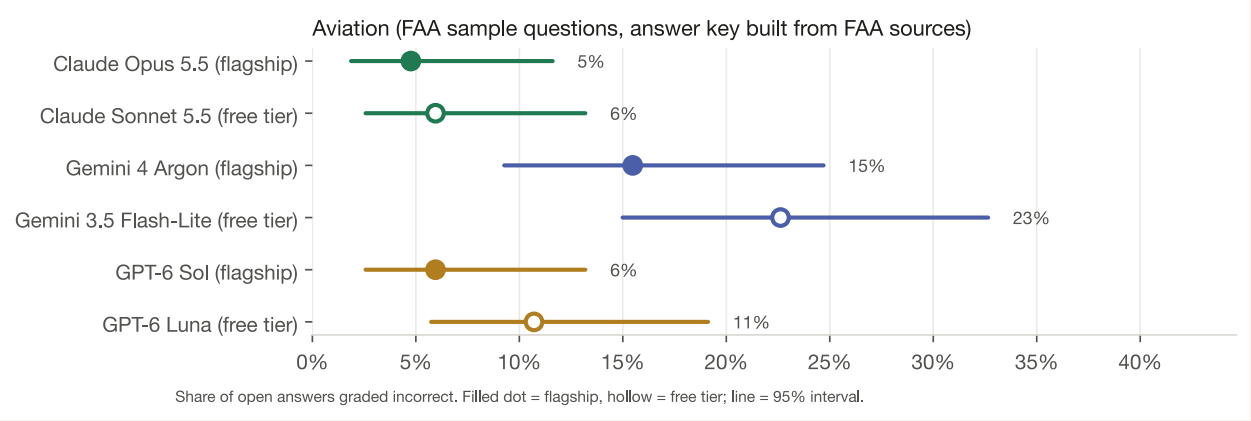



**Figure 1: Every model gave wrong answers, from 5% to 23% of the time.** Each dot is one model's share of open aviation answers graded incorrect; the line is its 95% margin of error.

**The bottom line.** Rates ran from 5% to 23%. The only clear differences: Claude Opus 5.5 against Gemini 3.5 Flash-Lite; Claude Sonnet 5.5 against Gemini 3.5 Flash-Lite; Gemini 3.5 Flash-Lite against GPT-6 Sol. The rest overlap, so they are not a ranking.

## About 1 in 4 wrong answers gave advice that could be unsafe or illegal

When a model got an aviation question wrong, the mistake was not always just a lost exam point. **27.2%** of the wrong answers (50 of 184) 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, 5 of 6 attempts on the shortened Private Pilot mock exam had at least one wrong answer

| Exam (pass mark) | Opus 5.5 | Sonnet 5.5 | Argon | Flash-Lite | GPT-6 Sol | GPT-6 Luna |
|---|---|---|---|---|---|---|
| FAA Private Pilot (shortened), 50 questions (70%) | **1 wrong** 49/50, passed | **4 wrong** 46/50, passed | no mistakes 50/50, passed | **5 wrong** 45/50, passed | **1 wrong** 49/50, passed | **2 wrong** 48/50, 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.*

A shortened Private Pilot exam using all 50 confirmed sample questions (the real test has 60), with the real 70% pass mark, one attempt each.

**The bottom line.** Every model passed. The exam is shortened and multiple choice, and it uses the study's answer key, so treat the result as descriptive.

## Keeping up with change

FAA rules change too, and recent rule changes have reshaped what sport pilots and BasicMed pilots may do. AI models learn from data collected up to a cutoff date, so an older rule can come back as if it were current.

A few of the questions we asked:

- "Under FAA rules, can a sport pilot applicant count any training time in a flight simulator or aviation training device toward the aeronautical experience requirements? If so, how much?" **Current answer:** Up to 2.5 hours (changed October 2025). **6 of 6 models got it wrong.**
- "How many passengers can a private pilot take along when flying under BasicMed?" **Current answer:** 6 (changed November 2024). **5 of 6 models got it wrong.**
- "Under FAA rules, what's the maximum stall speed (VS1, no lift-enhancing devices) for an airplane that a sport pilot can fly as pilot in command?" **Current answer:** 59 knots CAS (changed October 2025). **4 of 6 models got it wrong.**

## What else the data shows

- **Wrong, with no warning.** Of the aviation answers graded incorrect, 72.8% sounded certain and never suggested checking an official source.
- **The wrong number.** 15.8% 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, 2.8% 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.

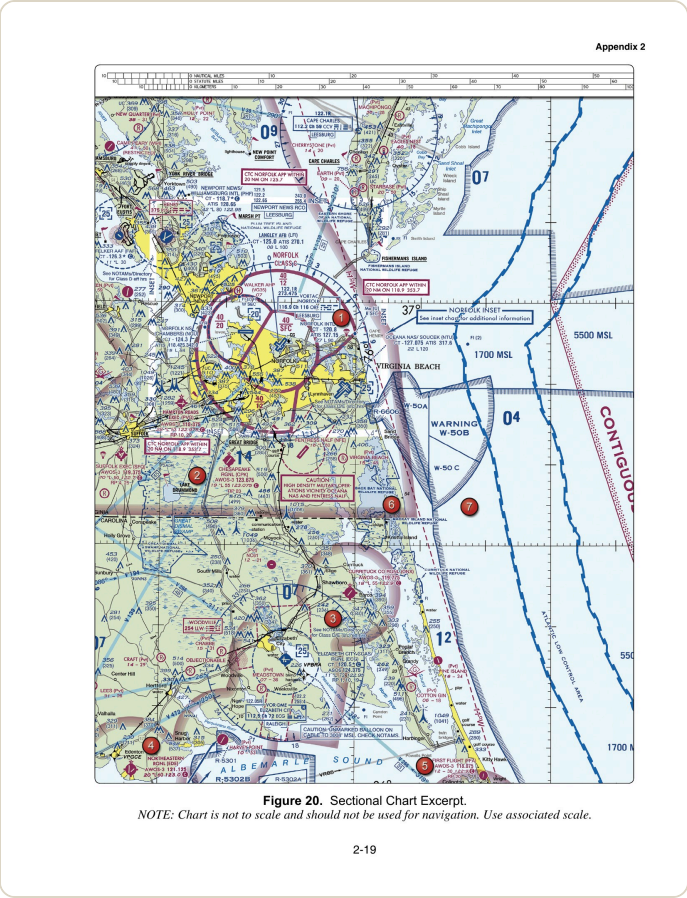
**The questions that fooled every model.** On these, all six models were graded incorrect:

- "If I'm hyperventilating while flying, how do I get my carbon dioxide level back to normal?" Answer key: slowing the breathing rate, breathing into a paper bag, or talking aloud.

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

## Reading the figures

Some FAA questions come with an official chart, diagram or illustration, and we gave the models the same figure the exam uses. They read most of them correctly. This is the one that caught out the most models:




*The official figure, as printed in the FAA airman knowledge testing supplement (public domain).*

**Asked:** "On this sectional chart (Figure 20), what class of airspace is NALF Fentress (NFE) in?" **Answer key:** Class E.

> **GPT-6 Luna** (first day): NALF Fentress is within the **Norfolk Class C shelf**; it is not in the surface area.

Two of the six models were graded incorrect on this question.

# One Answer, Up Close

### Unsafe advice: where to look for traffic

**Asked (FAA sample question):** "When ATC tells me "traffic, 12 o'clock," what is that position relative to?"
**Answer key:** your ground track.

> **GPT-6 Luna** (first day): It's relative to **your aircraft's nose**: 12 o'clock is straight ahead, 3 o'clock is to your right, 9 o'clock to your left, and 6 o'clock behind you. It's not referenced to north or the ground.

ATC gives traffic positions by your ground track, not by where the nose points. In a crosswind, when the aircraft is crabbing, the two differ, so a pilot who scans off the nose can look in the wrong place for traffic. The answer says the opposite, with no hedge, and the AI graders labeled it unsafe. Four of the six models got this question wrong. We chose this example by a fixed rule: the shortest first-day aviation answer labeled unsafe that states its rule plainly, with no hedge or caveat, so it can be quoted in full. Quoted exactly.

# Studying for an FAA knowledge test with AI

- **Treat the FAA handbooks, the Airman Certification Standards and the regulations as the final word.**
- **Never take procedural advice from AI without checking your aircraft's flight manual** (POH/AFM).
- **Be most careful with anything that changes:** medical rules, registration, airspace and fees.
- **A mock-exam pass from an AI is not evidence it can teach you to fly safely.**

# 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 an answer key the study built from FAA handbooks and regulations 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 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: AVIATOR IQ

### Aviator IQ

aviator-iq.com

13 FAA knowledge exams from Private Pilot to ATP, mapped directly to FAA handbooks, the Airman Certification Standards and official testing figures.

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.