oddly sunny

When AI Fails at Ham Radio Exam Prep

Six leading AI models from Anthropic, Google and OpenAI all pass the FCC Technician mock exam. Ask the same official question-pool items without the options, the way people actually study, and they get about seven times as many wrong against the official answer key.

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 NCVEC question-pool items (Technician, General and Extra) with the official answer key.

Summary

Amateur radio exams draw on public question pools, so you might expect AI models to know them cold. In multiple choice they nearly did. Asked the same pool questions as open questions, the models gave many more answers graded incorrect; the example below shows one way, the right idea mixed with one of the pool's wrong options. Five of the official questions involve rules that changed in February 2026, and our separate questions on recently changed rules, written from current government sources (where errors were highest overall), included 10 on radio; both are too few to report on their own.

The multiple-choice scores, and a perfect or near-perfect Technician mock exam for every model, give a false sense of security about how a model answers a candidate's open question.

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 2.2% to 15.3%

Format	Questions	Graded incorrect
Official multiple choice	85	2.2% (1.2% to 3.8%)
Open question, no options	85	15.3% (12.4% to 18.7%)

The same amateur radio 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 2.2% to 15.3%. 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 7% to 33% of the time

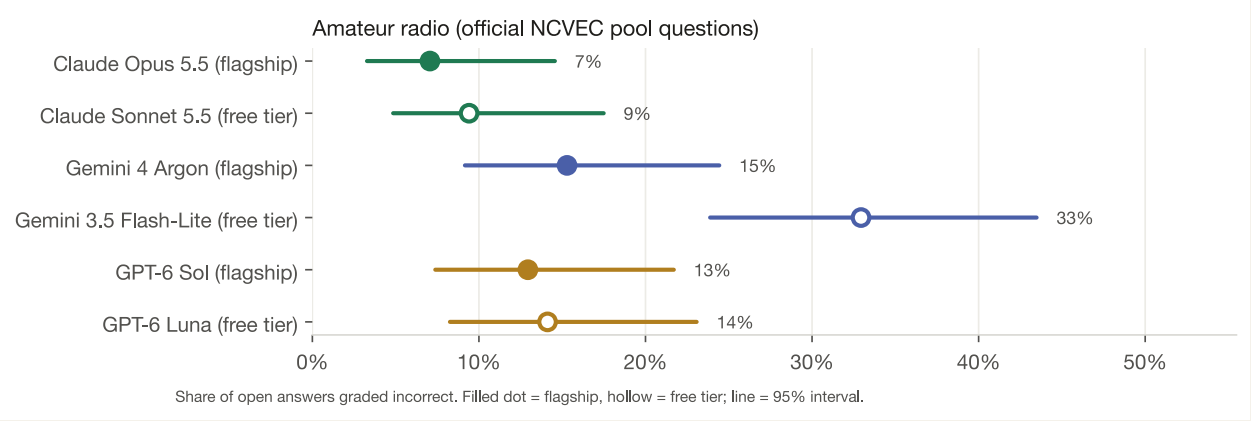


Figure 1: Every model gave wrong answers, from 7% to 33% of the time. Each dot is one model's share of open amateur radio answers graded incorrect; the line is its 95% margin of error.

The bottom line. Rates ran from 7% to 33%. 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; Gemini 3.5 Flash-Lite against GPT-6 Luna. The rest overlap, so they are not a ranking.

About 1 in 13 wrong answers gave advice that could be unsafe or illegal

When a model got an amateur radio question wrong, the mistake was not always just a lost exam point. **7.5%** of the wrong answers (19 of 253) 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 1.2%.

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, 3 of 6 attempts on the Technician 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
FCC Technician, 35 questions (26 of 35)	**1 wrong** 34/35, passed	no mistakes 35/35, passed	no mistakes 35/35, passed	**1 wrong** 34/35, passed	no mistakes 35/35, passed	**1 wrong** 34/35, 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 full-length Technician exam (35 questions) built from the official pool, in multiple-choice format, one attempt each.

The bottom line. Every model passed easily. In our view, that makes a mock-exam score a poor guide to how well a model explains the material.

Keeping up with change

FCC amateur rules change too, including recent changes to the 60-meter band and to equipment rules. 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:

- "Where in the 60-meter band can a General class amateur station transmit?" **Current answer:** In the contiguous segment 5351.5-5366.5 kHz and on four discrete channels (5332, 5348, 5373 and 5405 kHz) (changed February 2026). **6 of 6 models got it wrong.**
- "Under FCC Part 97, is an external RF power amplifier (capable of operation below 144 MHz) exempt from the certification requirement because an amateur bought it before April 28, 1978?" **Current answer:** No; that pre-1978 exception has been removed. The remaining exceptions cover amplifiers built or modified by an amateur for amateur use and certain sales to amateurs or dealers (changed February 2026). **6 of 6 models got it wrong.**
- "Can a dealer sell an external RF power amplifier that works below 144 MHz if it hasn't been FCC certified? When?" **Current answer:** The amplifier is constructed or modified by an amateur radio operator for use at an amateur station (changed February 2026). **5 of 6 models got it wrong.**

What else the data shows

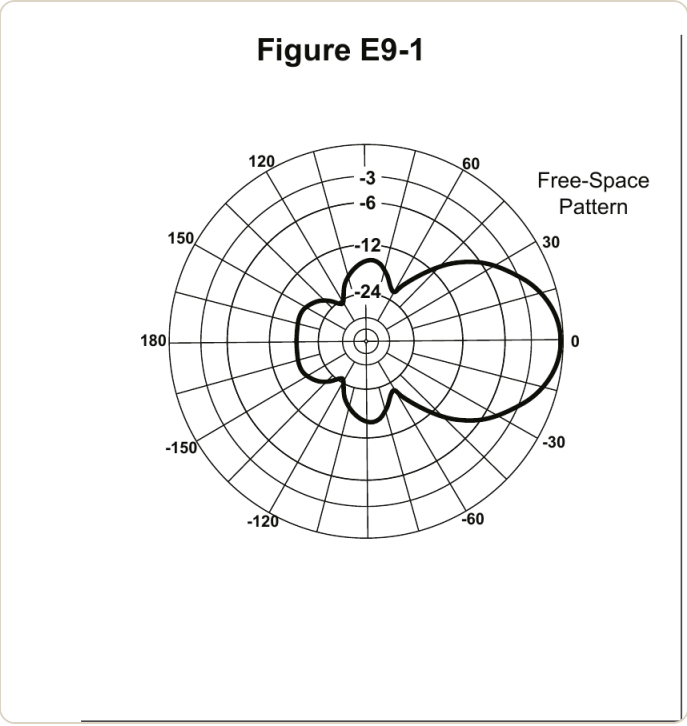
- **Wrong, with no warning.** Of the ham radio answers graded incorrect, 86.6% sounded certain and never suggested checking an official source.
- **The wrong number.** 14.2% of the wrong answers got a figure wrong: a distance, a limit, a fee or a percentage.

- **Right answer, shaky reasoning.** Even among answers graded correct, 3.1% 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.

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

Reading the figures

Some question-pool items 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 FCC amateur radio question pool (public domain).

Asked: "What's the 3 dB beamwidth of the antenna pattern in this figure?" **Answer key:** 50 degrees.

> **GPT-6 Luna** (first day): The main lobe crosses the −3 dB ring at about +45° and −45°, so the 3 dB beamwidth is approximately **90°**.

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

One Answer, Up Close

The right idea plus a wrong option: a power supply interlock

Asked: "What's the purpose of a power supply interlock?" **Official answer (General pool, G0B12):** to ensure that dangerous voltages are removed if the cabinet is opened.

> **GPT-6 Sol** (first day): A power supply interlock prevents a supply from turning on—or shuts it off—unless required safety conditions are met, such as a cover being closed or cooling being available. It protects people and equipment from hazardous power.

The answer mentions the cover being closed, but it also lists cooling, which matches the pool's wrong option "to shut down the unit if it becomes too hot", so it was scored incorrect. Five of the six models were graded incorrect on this question on at least two of three days. Quoted exactly.

Studying for a ham radio exam with AI

- **Study from the current official pool.** Check which pool is in force for your exam date; the pools change on a fixed cycle.
- **Use AI to explain the theory, then check the pool's keyed answer** before you memorize anything.
- **Watch for answers that list several purposes or causes.** Each question keys one answer, and the extras can match the distractors.
- **Check rule questions against Part 97 itself,** especially anything about privileges, power limits or identification.

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 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 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: AIRWAVES IQ

Airwaves IQ

airwaves-iq.com

FCC Technician, General, Extra, GROL and GMDSS exams, synchronized to the current NCVEC question pools and FCC Part 97.

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.