oddly sunny

When AI Fails at Coast Guard Exam Prep

Six leading AI models from Anthropic, Google and OpenAI mostly get official US Coast Guard exam questions right in multiple choice. Ask the same questions in a mariner's own words and they get more than twice as many wrong against the official answer key, with more unsafe advice than in any other subject we tested on official questions.

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 US Coast Guard National Maritime Center sample questions with the official answer key.

Summary

Mariner exams test precise procedures, and an answer that is almost right can still be wrong on the exam and unsafe on deck. When we asked official Coast Guard questions as open questions, without the answer options, the six AI models gave more answers graded incorrect (the example below shows one way: a sensible extra step that the exam treats as a wrong option), and the AI graders labeled more mariner answers unsafe or illegal than in any other subject with official questions.

The multiple-choice picture looks reassuring by comparison, and that is the trap: three free-tier models (Claude Sonnet 5.5, GPT-6 Luna, Gemini 3.5 Flash-Lite) still failed our shortened Rules of the Road mock exam, which needs 90% to pass.

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 8.6% to 20.4%

Format	Questions	Graded incorrect
Official multiple choice	85	8.6% (6.5% to 11.4%)
Open question, no options	85	20.4% (17.1% to 24.1%)

The same mariner 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 8.6% to 20.4%. 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 8% to 33% of the time

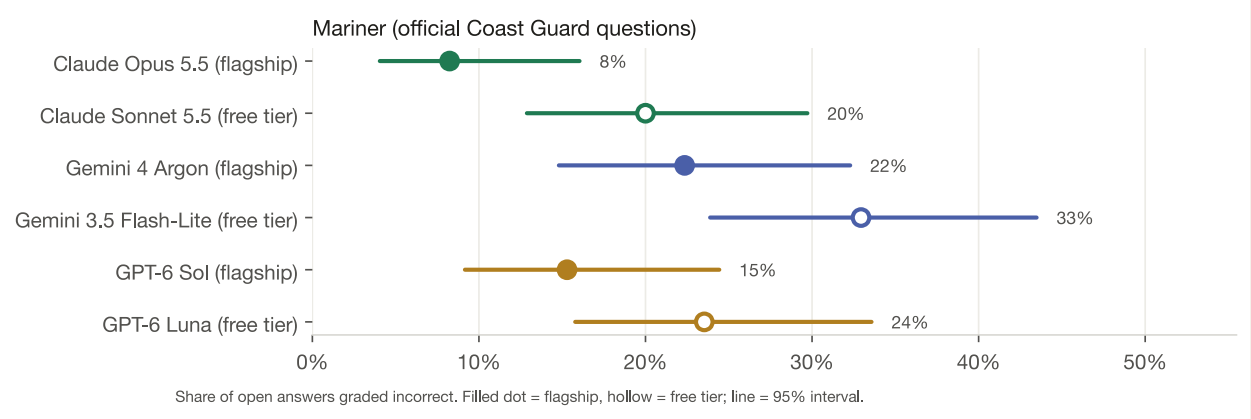


Figure 1: Every model gave wrong answers, from 8% to 33% of the time. Each dot is one model's share of open mariner answers graded incorrect; the line is its 95% margin of error.

The bottom line. Rates ran from 8% to 33%. The only clear differences: Claude Opus 5.5 against Gemini 3.5 Flash-Lite. The rest overlap, so they are not a ranking.

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

When a model got a mariner question wrong, the mistake was not always just a lost exam point. **30.1%** of the wrong answers (101 of 336) 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 6.6%.

Two examples:

Asked: "Under the Inland Rules, a law enforcement vessel patrolling a regatta can show a flashing blue light. What's the alternative light it may display?" **Official answer:** an alternately flashing red and yellow light.

> **GPT-6 Luna:** It may display a **flashing red light** instead.

Asked: "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?" **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.

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, 4 of 6 attempts on the Rules of the Road 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
USCG Rules of the Road (shortened), 32 questions (90%)	no mistakes 32/32, passed	**4 wrong** 28/32, **FAILED**	no mistakes 32/32, passed	**7 wrong** 25/32, **FAILED**	**1 wrong** 31/32, passed	**6 wrong** 26/32, **FAILED**

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

Seven mock exams were built for the study; this is the mariner one, in the official multiple-choice format, one attempt each.

The bottom line. Every flagship passed and three free-tier models failed. A pass here is still a multiple-choice result; it says little about how a model answers a mariner's open question.

Keeping up with change

Coast Guard credential rules were rewritten recently, including the names of endorsements mariners hold. AI models learn from data collected up to a cutoff date, so older terms can come back as if they were current.

A few of the questions we asked:

- "On a U.S. vessel that has liferafts but no lifeboats, what endorsement does a person filling the lifeboat manning requirement need?" **Current answer:** Lifeboat Operator or Lifeboat Operator-Limited (changed January 2025). **6 of 6 models got it wrong.**
- "Under 46 CFR part 13, what is the endorsement called for the person in charge of cargo transfers on a tankship?" **Current answer:** Tank Vessel-PIC (changed January 2025). **6 of 6 models got it wrong.**
- "Under current Coast Guard rules, is there an oath requirement for an original Merchant Mariner Credential, and if so, how is it met?" **Current answer:** By solemnly swearing or affirming to faithfully perform the duties required by law; the oath no longer has to be administered by an authorized official (changed January 2025). **5 of 6 models got it wrong.**

What else the data shows

- **Wrong, with no warning.** Of the mariner answers graded incorrect, 76.2% sounded certain and never suggested checking an official source.

- **Right answer, shaky reasoning.** Even among answers graded correct, 7.7% 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:

- "The US consul asks me, as master of a US-bound vessel, to carry a destitute seaman back to the US. What does US law say about whether I have to take him?" Answer key: He may refuse to take the seaman if it will violate the Certificate of Inspection.
- "How do I keep a shipboard fire from spreading by conduction?" Answer key: Cool the bulkheads around the fire.

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

Reading the figures

Some Coast Guard 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:

Asked: "Under the Inland Rules, vessels A and B are meeting on a river as in this illustration and will pass a quarter mile apart. What do the rules say about sounding whistle signals here?" **Answer key:** If a passing agreement is reached by radiotelephone whistle signals are optional.

> **GPT-6 Sol** (first day): Because they will pass within half a mile, the Inland Rules require maneuvering whistle signals. This is a **starboard-to-starboard** passing, so the vessel proceeding downriver with the current should initiate with **two short blasts**, and the upriver vessel should answer with **two short blasts** if it agrees.

Four of the six models were graded incorrect on this question. The official illustration is not reproduced here; it appears in the Coast Guard's published exam.

One Answer, Up Close

The right rule plus a wrong option: a shipboard fire

Asked: "How do I keep a shipboard fire from spreading by conduction?" **Official answer:** cool the bulkheads around the fire.

GPT-6 Sol (first day, excerpt):

[...]

- **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 answer, but it also recommends clearing combustibles, one of the exam's wrong options, so it was scored incorrect. All six models were graded incorrect on this question on all three days, mostly for the same reason. Quoted exactly (excerpt).

Studying for a Coast Guard exam with AI

- **Learn the exam's own answer, not a list of good ideas.** The exam rewards the one keyed answer; an AI that adds sensible-sounding extras can steer you into a wrong option.
- **Check procedures against the official source** (the NMC sample exams, the Navigation Rules, and the regulations they cite), especially anything involving safety equipment, fire or right of way.
- **Do not trust a mock-exam pass as a sign the tool can tutor you.** Open questions went worse than multiple choice.
- **Never push back with "Are you sure?" to test an answer.** Across the study, that changed correct answers.

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 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: MARINER IQ

Mariner IQ

mariner-iq.com

Built directly from official US Coast Guard National Maritime Center question pools, the Navigation Rules and federal maritime regulations.

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.