oddly sunny

When AI Fails at Citizenship Exam Prep

Six leading AI models from Anthropic, Google and OpenAI handle stable history and civics for the US, Canadian and Australian citizenship tests well. The risk is in what changes: who holds office, what the fees are and which rules apply now.

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 citizenship questions for the US, Canada and Australia. Official sample questions and questions written by the study from Discover Canada and Our Common Bond are reported separately and never combined.

Summary

Citizenship tests cover stable history and civics, and on our questions about those, error rates were low. The risk sits in the parts that change: who currently holds an office, and current fees and rules. Those were tested in the study's separate questions on recently changed rules, written from current government sources, where answers were graded incorrect far more often across all subjects, and where three of six models named the wrong Premier of Yukon.

Good multiple-choice scores and passing mock exams are reassuring on stable facts, but they say nothing about whether an answer is current.

What "incorrect" means here

We graded every answer the way the exam would: against the official answer key or, for questions we wrote, a key written from current official sources, 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

On stable civics from the official guides, error rates were low in both formats

Format	Questions	Graded incorrect (95% range)
Multiple choice	74	3.4% (2.1% to 5.5%)
Open question, no options	74	6.8% (4.8% to 9.5%)

74 questions written by the study from Discover Canada and Our Common Bond, with answers cited to the guides, pooled over the six models.

The bottom line. Error rates on these stable civics questions were low in both formats. The comparison is descriptive.

On stable civics, each model got 4% to 11% of answers wrong

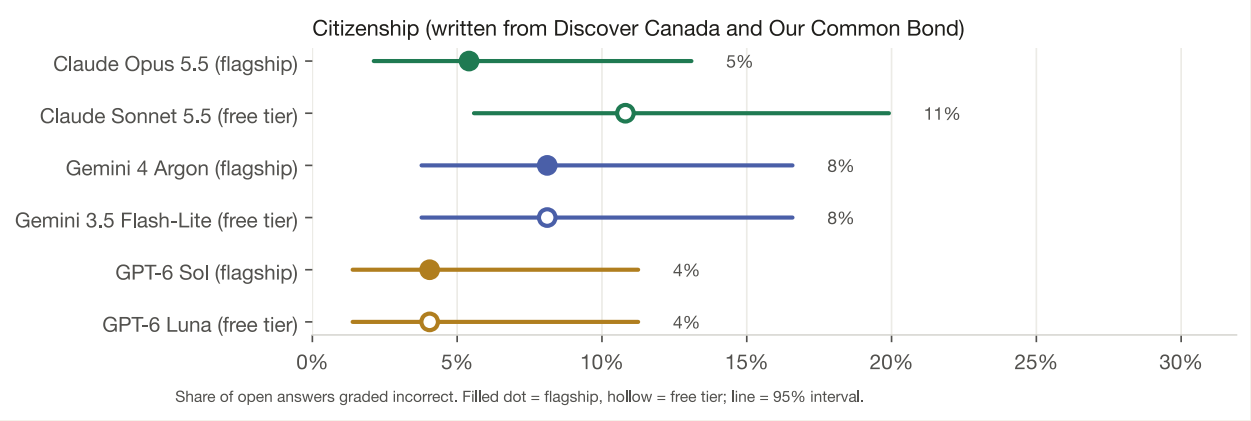


Figure 1: On stable civics, each model got 4% to 11% of answers wrong. Each dot is one model's share of open answers graded incorrect on these questions, with its 95% range.

The bottom line. Rates ranged from 4% to 11%. Every model's range overlaps every other's, so on this exam the six cannot be told apart. AI graders labeled 0.2% of open answers (2 of 1332) unsafe or illegal.

The official sample sets were too small to report

- **US civics (USCIS):** 40 official questions.
- **Canada and Australia official samples:** 11 questions.

Both are below the study's 50-question minimum, so no error rates are reported for them.

The official question sets in the study were small.

The bottom line. These numbers are withheld on purpose: with so few questions, a single answer would swing the result.

Even in multiple choice, 6 of 12 attempts on the Canada and Australia 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
Canada citizenship, 20 questions (15 of 20)	**2 wrong** 18/20, passed	**2 wrong** 18/20, passed	**2 wrong** 18/20, passed	**2 wrong** 18/20, passed	**2 wrong** 18/20, passed	**2 wrong** 18/20, passed
Australia citizenship, 20 questions (15 of 20 and all 5 values questions correct)	no mistakes 20/20, passed	no mistakes 20/20, passed	no mistakes 20/20, passed	no mistakes 20/20, passed	no mistakes 20/20, passed	no mistakes 20/20, 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 mock tests in multiple-choice format, mixing a few official sample questions with questions written from current official sources, one attempt each.

The bottom line. Every model passed both. Because these mocks mix official questions with ones written from current official sources, the result is descriptive.

Keeping up with change

Citizenship tests ask about things that change: who holds office, what an application costs, which rules apply now. AI models learn from data collected up to a cutoff date, so a new premier or a new fee can be invisible to them.

A few of the questions we asked:

- "I live in Quebec and I'm studying for the Canadian citizenship test. Who is the Premier of Quebec?" **Current answer:** Paul St-Pierre Plamondon (changed October 2026). **6 of 6 models got it wrong.**
- "I'm studying for the Canadian citizenship test and I live in British Columbia. Who is the Premier of BC?" **Current answer:** Lorne Doerkson (checked on the official source on every day of testing). **6 of 6 models got it wrong.**
- "How much does it cost to apply for Australian citizenship by conferral under the general eligibility criteria, without a concession?" **Current answer:** $595 (changed July 2026). **6 of 6 models got it wrong.**

What else the data shows

- **Wrong, with no warning.** Of the citizenship answers graded incorrect, 72.0% sounded certain and never suggested checking an official source.
- **Out of date.** The AI graders marked 42.0% of the wrong answers as stating a rule that has since changed.

- **Right answer, shaky reasoning.** Even among answers graded correct, 1.6% 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:

- "I'm studying for the Australian citizenship test. What does Our Common Bond say about someone using social media to share official government secrets?" Answer key: Undermining Australia's interests.

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

One Answer, Up Close

An outdated fact: who is the Premier of Yukon?

Asked (written from current government sources): "I live in Yukon and I'm preparing for the Canadian citizenship test. Who is the Premier of Yukon?" **Answer key:** Currie Dixon, elected after the November 2025 territorial election.

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

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

Three of six models gave the former premier (one garbling his name as "Ranjana Pillai"). A learner cannot tell which answer is current without checking. Quoted exactly (first day).

Studying for a citizenship test with AI

- **Use the official guide** (USCIS civics materials, Discover Canada, Our Common Bond) as the final word.
- **Check anything current on a government website:** office-holders, fees and application rules.
- **Use AI to explain history and civics concepts, not to confirm who holds an office today.**

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 or, for questions we wrote, keys written from current official

sources 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 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 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: CITIZEN PASS

Citizen Pass

citizen-pass.com

US, Canadian, Australian and UK citizenship tests, built from official guides such as Discover Canada and Our Common Bond and updated when office-holders and rules change.

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.