

The Xzar Report: Frontier AI on real patent decisions

Published October 2026 · Xzar, the crash-test lab for AI

Summary

- On 60 real patent examination decisions, all made after the model was trained, a frontier AI scored **0.12** out of 1. A one-line rule of thumb scored **0.36** and beat the AI on **47 of the 60** cases.
- Given search over prior art and the examination rulebook, the same AI reached **0.22**, still below the rule of thumb.
- The AI's mistakes were lopsided: when it was wrong, **92%** of the time it let through a claim the examiner had rejected.

Why test AI on real decisions

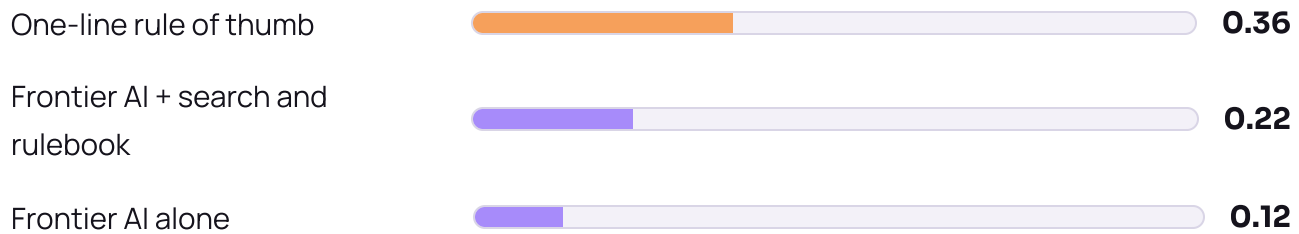
AI agents are being deployed to do professional work: reviewing claims, drafting decisions, checking compliance. Most benchmarks test knowledge with questions written for the test. We test the work itself. Each task is a real case that a professional actually decided, and the AI is graded against what was actually decided.

How a test works

1. **A real case, cut at the decision.** The AI receives what the examiner had at that moment: the application's claims and description as filed, a library of prior publications dated before the filing date, and the examination rulebook. Nothing from after the decision is included.
2. **The AI works the case.** It can search the prior-art library and the rulebook, then decides, for each claim, whether to reject it, on which ground (for example novelty or obviousness), and citing which prior art.
3. **Exact grading against what happened.** Each answer is scored from 0 to 1 against the examiner's actual action: the decision on each claim, the grounds, and the evidence cited. Grading is deterministic: the same answer always gets the same score. No AI judges.

4. **Cases the model can't have memorized.** Every case in this report was decided after the model's training cutoff. Answer keys never leave our servers.

Results



- **Frontier AI alone: 0.12** (95% range 0.09 to 0.16), on 60 cases.
- **One-line rule of thumb: 0.36** (95% range 0.30 to 0.41) on the same 60 cases. The rule: reject every claim as obvious. It beat the AI on 47 of 60.
- **Frontier AI with search and the rulebook: 0.22**, on the 59 matched cases where a library was available. Tools helped, but not enough to reach the rule.
- **The failure pattern:** of 506 wrong calls on individual claims, 466 (92%) were claims the AI let through that the examiner rejected. An AI deployed to clear claims would mostly fail in the risky direction.

How hard this is for people

Real professional decisions are not easy, and experts disagree. That is the standard AI is being held to.

- **Examiners differ.** Within the same examining unit, the grant rates of examiners differ by a median of 37 percentage points (the gap between the 10th and 90th percentile examiner, across 592 examiners with at least 20 resolved applications).
- **Contested decisions are close.** When a rejection is appealed, the appeal board agrees with the examiner about 65% of the time (9,083 appealed claims).

A good AI doesn't need to be perfect. It needs to beat simple rules, fail in known directions, and know when to hand a case to a person. Today it does none of these reliably on this task.

Limitations

- **Sample size.** 60 cases is a first public cut. The 95% ranges are reported, and the gap between the AI and the rule of thumb holds across them.
- **One model family.** Results are for one frontier model family. Our API lets anyone test other models on the same cases.
- **One decision type.** These are first examination decisions. Our other live tests (the decision after the applicant responds, and appeals) are reported separately.
- **Search coverage.** The prior-art libraries contain about two thirds of the art the examiners actually cited, so the tool-assisted score is a floor for what better search could achieve.
- **Human comparisons are coarse.** Grant-rate spread and appeal agreement show how much experts differ; they are not a direct accuracy score for examiners.

Method notes

Scores are means across cases; 95% ranges come from 5,000 bootstrap resamples. The rules of thumb are scored on exactly the same cases as the AI. The grader is versioned, and every graded answer gets a receipt listing what matched and what didn't.

Test your own AI

- **Self-serve API:** an API key with \$5 of free credit at <https://xzar.si/signup.html>, then \$2.00 per scored task. Or add the MCP plug-in: `uvx xzar-mcp`.
- **Pilot:** 30 days on our cases with a full scorecard and a written failure analysis.
- **Your field:** patents are the first test. If professionals decide it, we can build a test for it.

Contact: hello@xzar.si · <https://xzar.si>