AI BUILT FOR LAW OUTPERFORMS CHATGPT, CLAUDE, AND GEMINI ON LEGAL REASONING BENCHMARK

DescrybeLM answered all 200 multistate bar exam questions correctly. ChatGPT, Claude, and Gemini each missed between 13 and 23 questions — and scored lower on legal reasoning quality across the board. Methodology and scoring rubric published.

BOSTON, MA — March 5, 2026

When AI gets a legal question wrong, the most dangerous failure isn't an obvious error. It's an answer that sounds authoritative: fluent, confident, well-structured, and yet applying the wrong legal standard. The error reads like competent lawyering.

Today, Descrybe launched DescrybeLM — an AI system built specifically for legal reasoning — and published a white paper with the benchmark data to show what that difference looks like in practice.

Descrybe ran a controlled benchmark against ChatGPT 5.2, Claude Opus 4.5, and Gemini 3 Pro on 200 multistate bar exam questions. The study measured not just whether each system chose the correct answer, but whether the legal reasoning behind it was sound: Did it identify the right rule? Apply it correctly to the facts? Avoid the traps that produce persuasive but wrong analysis?

"We had a thesis that purpose-built legal AI produces meaningfully different results for legal reasoning tasks. Legal professionals deserve to make tool decisions based on real evidence, which can be hard to find. So, we tested ourselves. We know that vendor-produced benchmarks invite scrutiny — as they should, that's why we published our methodology and invite anyone to replicate it," said Kara Peterson, Co-Founder and CEO of Descrybe.

What the benchmark showed

All four systems were tested under standardized, no-external-web benchmark conditions using the NCBE MBE Complete Practice Exam (Questions 1–200, no exclusions), producing 800 separate evaluation runs with blinded scoring.

System	Correct Answers out of 200	Rubric-Scored Reasoning Quality
DescrybeLM	200 (100%)	99.70%
ChatGPT 5.2	187 (93.5%)	93.41%
Claude Opus 4.5	177 (88.5%)	89.03%
Gemini 3 Pro	184 (92.0%)	91.45%

Rubric-scored reasoning quality reflects judge-model scoring; see Evaluation Approach within the white paper for methodology and limitations. As with any benchmark using a commercially available question set, training data exposure cannot be ruled out for any evaluated system, including DescrybeLM.

When general-purpose models were wrong, they were confidently wrong. Among 52 incorrect outputs, 49 delivered assertive, well-structured reasoning that did not signal uncertainty — the failure mode that imposes the highest verification burden on legal practitioners. The dominant patterns were applying the wrong legal standard or misapplying the correct one, while the prose read like competent analysis.

The scoring log also revealed that two general-purpose models — Claude Opus 4.5 and Gemini 3 Pro — exhibited overconfident tone on correct outputs, not only incorrect ones. DescrybeLM and ChatGPT 5.2 each received zero overconfidence flags across all 200 outputs. A system that sounds equally confident whether it is right or wrong gives practitioners no reliable signal from tone alone.

The study also found that cross-checking between general-purpose models is not a reliable substitute for getting the answer right the first time. Across 200 questions, 40 were missed by at least one model, 11 were missed by two or more, and only 1 was missed by all three — meaning errors were largely unpredictable and non-overlapping.

What's behind the results

DescrybeLM is built on a curated primary law corpus of more than 100 million structured records, cleaned and processed at a scale requiring more than 100 billion tokens of preparation. The system is optimized for reliability and verification.

“Most AI tools are built for general use and adapted for law. DescrybeLM was built differently: from the foundation up, specifically for legal reasoning, on more than 100 million structured records individually cleaned and organized for that purpose. That kind of data work is painstaking and takes years, but it's the difference between a system that sounds right and one that is right,” said Richard DiBona, Co-Founder and CTO of Descrybe.

Why this matters

The headline problem in legal AI isn't systems that obviously fail. It's systems that fail invisibly, confidently, and in a way that reads like competent analysis. And in a crowded market, sounding right is easy to mistake for being right. Legal professionals need real evidence to decide which tools they should use for which purposes. That's why Descrybe published this benchmark's methodology and invites independent replication.

“I've worked in legal technology for a long time. It's rare to see something that genuinely stops you in your tracks. When I saw DescrybeLM answer all 200 multistate bar exam questions correctly while ChatGPT, Claude, and Gemini each missed double digits, that's exactly what happened. That's not a marginal difference — that's a different category of tool,” said Ken Friedman, legal technology pioneer and advisor to Descrybe.

The full white paper, *Beyond Confidently Wrong: How Purpose-Built AI Mitigates Legal Reasoning's Hidden Risk*, is available here.

About Descrybe

Descrybe builds AI designed specifically for legal reasoning, grounded in a structured primary-law corpus of more than 100 million records. For more information, visit descrybe.com.

Media Contact: Kara Peterson | kara@descrybe.com | 617-752-2020

###