Beyond Confidently Wrong: How Purpose-Built AI Mitigates Legal Reasoning's Hidden Risk

Executive Summary

In legal work, the most dangerous AI failure is not an obvious error or hallucination. It is an answer that is wrong but sounds authoritative. A fluent, confident, incorrect output can waste time, misdirect strategy, and create downstream risk that is difficult to detect before damage is done.

This paper reports a controlled benchmark comparing DescrybeLM™, an AI-native, purpose-built legal reasoning system, against three leading general-purpose models (ChatGPT 5.2, Claude Opus 4.5, and Gemini 3 Pro) on the National Conference of Bar Examiners ("NCBE") Multistate Bar Examination ("MBE") Complete Practice Exam. The test comprised 200 multiple-choice bar exam questions scored under standardized, no-external-web conditions (no public web browsing; each system relied only on its own internal knowledge or data) with blinded evaluation. Each system was asked not only to select the correct answer but to explain its reasoning — identifying the governing legal rule, applying it to the facts given, and addressing why each other option was wrong. Both dimensions were scored under the same rubric across all four systems.

The results were striking: DescrybeLM answered all 200 questions correctly. The general-purpose models each missed between 13 and 23 questions (88.5%–93.5% accuracy), with rubric-scored reasoning quality ranging from 89.0% to 93.4% compared to 99.7% for DescrybeLM.

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%

Correct/incorrect results are independently verifiable using the commercially available answer key. Rubric-scored reasoning quality reflects judge-model scoring; see Evaluation Approach for methodology and limitations.

The more revealing finding is how errors were distributed. Across the three general-purpose models, 40 unique questions were answered incorrectly by at least one system, but only 1 question was missed by all

three. The remaining 39 were missed by one or two models, meaning no single model's errors predict another's. In practice, this means cross-checking between general-purpose models is not a reliable substitute for getting the answer right the first time.

When these systems were wrong, they were confidently wrong. Among the 52 total incorrect outputs, the dominant failure patterns applied the wrong legal standard or misapplied the correct one, while presenting the analysis in fluent, well-structured prose. These are precisely the errors that impose the highest verification burden on practitioners.

This study has clear limits: it covers one benchmark format, it was conducted by the team that built DescrybeLM, and a perfect score on a commercially available question set invites reasonable scrutiny. One limitation applies equally to all systems tested: because the NCBE MBE Complete Practice Exam is commercially available, we cannot rule out that some or all items appeared in the training data of any evaluated system, including DescrybeLM. Fine-tuning performed on DescrybeLM prior to this benchmark was conducted on a separate NCBE product and did not use the question set or answer key evaluated here. Each question was presented as a first-seen input under standardized conditions. The paper documents these constraints in detail and invites independent replication. What the results do show, within that scope, is that purpose-built legal systems and general-purpose models can produce meaningfully different reliability profiles, and that the gap is widest precisely where this paper began: outputs that are wrong but sound authoritative.

Why This Study

Legal analysis is high-stakes, and time is scarce. When a persuasive but wrong answer enters a workflow, the downstream consequences can be severe. General-purpose LLMs are often excellent at producing fluent legal-style writing, and they are typically optimized to be helpful and responsive. But that same “helpfulness” creates a known risk pattern: when a model hits a gap in knowledge or context, it may try to fill the gap with plausible-sounding information rather than clearly indicating uncertainty.

This paper reports a comparative evaluation designed to test a thesis: purpose-built legal systems with access to clean and structured data may produce different accuracy and reasoning-quality profiles than general-purpose LLMs on standardized legal reasoning tasks. Descrybe 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 — a level of data preparation rarely undertaken in the legal AI space. Under consistent constraints, we measure how reliably different systems select the correct legal answer and explain it using defensible legal reasoning, and we report the resulting outcomes and observable failure modes rather than marketing claims.

How to Read This Paper

What this evaluation can and cannot tell you

This study uses an MBE-style multiple-choice benchmark as a proxy for assessing legal reasoning under controlled constraints. Multiple-choice questions do not capture every dimension of real legal work, but they do provide a standardized way to measure whether a system can (a) identify the governing rule, (b) apply it to the facts given, and (c) avoid high-friction failure modes like confidently wrong reasoning. This evaluation measured both the correct/incorrect answer choice and the legal reasoning why the choice was made as a proxy for legal reasoning strength.

This evaluation can tell you

- How systems performed on a controlled, exam-style multiple-choice benchmark.
- How often systems selected the correct option and the quality of their legal reasoning explanation (i.e., whether the output identifies the governing rule and applies it correctly to the facts given), as a proxy for legal reasoning under controlled constraints.
- Common failure modes that produce persuasive but incorrect outputs.

This evaluation cannot tell you

- How any system will perform on legal workflows beyond exam-style multiple-choice reasoning (e.g., drafting, negotiation, fact investigation, jurisdiction-specific research).
- Whether an output is current law for a specific jurisdiction without verification.

The reliability lens

When evaluating legal AI, we focus on five practical reliability questions:

1. Correctness: Is the answer right under the defined constraints?
2. Verification burden: How much work does a lawyer need to do to trust it? In practice, confidently wrong is the worst case here: a fluent but wrong answer can be harder to detect than an obviously uncertain response, increasing the time and expertise required to verify outputs.
3. Failure severity: When it's wrong, is it obviously wrong, or confidently persuasive?
4. Traceability: Can you quickly see and validate what the answer depends on?
5. Reasoning quality: Even when a system selects the correct option, did it do so for the right reasons (rule identification and correct application), in a way that would generalize to similar fact patterns?

This benchmark directly tests dimensions 1, 2, 3, and 5. Dimension 4, traceability, is central to real-world legal AI evaluation but is not directly tested in this no-external-web benchmark, which does not involve

source citation or authority retrieval. Future evaluations involving research-grounded tasks (Future Work) will assess traceability directly.

Systems Evaluated

DescrybeLM: An AI-native, purpose-built legal reasoning system designed to help users move from a legal question to a research-grounded answer. It is built on the Descrybe structured primary-law corpus and is optimized for reliability and verification. For this benchmark, DescrybeLM reflects a fine-tuned configuration intended to improve legal reasoning and reliability under no-external-web constraints (relying on its internal primary-law corpus, not the public web).

What it is (high level)

- A legal reasoning system that emphasizes accuracy, traceability, and risk control.
- Designed to surface key issues and apply generally accepted legal principles to the facts presented.

What it does (in this benchmark)

- Produces a single best-answer selection (A/B/C/D) under no-external-web constraints (relies only on its own corpus of primary law).
- Provides an explanation intended to be verification-friendly (clear rule statement + application to key facts).

We do not disclose the DescrybeLM base model, training data, or retrieval architecture in this paper for competitive reasons. The results should be evaluated based on the documented test conditions and outputs, not on architectural claims. This paper does disclose: the exact question set and selection rule (Study Design), the verbatim standardized prompt (Appendix C), all run conditions including interface, session control, and tool settings (Study Design), the full scoring rubric and weights (Appendix A), the blinding and scoring process (Evaluation Approach), and the per-output scoring log (Appendix B).

What it is not

- Not a web-browsing tool in this evaluation; DescrybeLM does not use the public web or external tools by default.
- The output is legal information only. Not a substitute for legal advice; outputs still require professional judgment and verification.
- General-purpose LLMs: ChatGPT 5.2, Claude Opus 4.5, and Gemini 3 Pro. All general-purpose models were accessed via their consumer interfaces or API during the January–February 2026 testing window, using paid versions available at the time of each run.

Model selection rationale

We selected widely used, commercially available model variants intended for day-to-day professional use. DescrybeLM, as an AI-native purpose-built system, was evaluated in the configuration used for product testing at the time of this benchmark (including fine-tuning). Providers offer multiple tiers and configurations; this evaluation reports results for the specific variants listed above.

This evaluation did not include other purpose-built legal research platforms. We invite other legal AI vendors to replicate this benchmark using the same commercially available question set, standardized prompt, and scoring methodology documented in this paper.

DescrybeLM includes an interactive mode that allows the system to ask follow-up and clarifying questions before generating a response. This mode was not used for this benchmark so that all systems were evaluated under the same single-pass, no-interaction constraints. The results reported here therefore reflect the DescrybeLM performance without one of its intended workflow features.

Disclosure

This evaluation was conducted and written by the team building DescrybeLM. We designed the study to be measurable and reproducible, with standardized instructions and blinded scoring, and we report limitations and observable failure modes alongside results.

A reasonable critic might also ask whether the rubric was designed, consciously or unconsciously, to favor DescrybeLM's output style. We took three steps to mitigate this risk. First, the rubric was authored by a human subject-matter expert and pre-committed before scoring began. Second, all outputs were anonymized before the judge model applied the rubric — the judge did not know which output belonged to which system. Third, the same rubric, judge prompt, and settings were applied identically across all 800 outputs. We cannot prove the absence of unconscious bias in rubric design, and we acknowledge this as a limitation. We invite independent researchers to apply their own rubric to the same outputs; the scoring methodology is fully documented to support that effort.

System-level comparison

This benchmark compares these systems as a user would actually use them, including whatever internal tools, data, or fine-tuning each product brings. DescrybeLM was evaluated in its product configuration (including internal primary-law grounding and fine-tuning), and the general-purpose models were evaluated as accessed via their standard interfaces. The results reflect how these systems performed under the same benchmark constraints.

Interpreting a 200/200 result

A perfect score can reasonably raise questions about data leakage, memorization, or other non-obvious advantages. This paper does not claim to prove the absence of any provider-side effects that are not

observable to evaluators. Instead, we report the controls we actually applied: a pre-committed non-cherry-picked question set (Q1–200 in order), standardized no-external-web instructions, no intentionally enabled browsing or external tools, consistent run conditions as documented, and blinded scoring applied uniformly across systems. Readers should interpret these results as a point-in-time measurement on this benchmark under these documented conditions, not as a guarantee of universal performance.

Because the NCBE MBE Complete Practice Exam is commercially available, we cannot rule out that some or all of these items appeared in the training data of any evaluated system, including DescrybeLM. This is an inherent limitation of benchmarking with published question sets and applies equally to all systems tested.

What we did (and did not do) to improve DescrybeLM performance

In development testing on a separate set of NCBE-authored MBE-style questions (a different NCBE product from the MBE Complete Practice Exam used in this benchmark), the DescrybeLM base configuration missed fewer than 5 out of 200 questions. Analysis of those misses indicated that the system was not reliably identifying the applicable area of law for certain question types. We fine-tuned the model to improve legal-domain recognition and routing, i.e., the system's ability to identify which area of law governs a question and apply the corresponding reasoning pathway. This fine-tuning was not performed on the NCBE benchmark items used in this evaluation, and the NCBE answer key was not provided to the system. Each benchmark question was presented as a first-seen input under the standardized instructions.

Why perfect accuracy does not imply perfect reasoning

DescrybeLM selected the correct option on all 200 questions, but 27 outputs received rubric scores below 100. The deductions were concentrated at the margin: 20 outputs scored 99, four scored 98, one scored 97, one scored 92, and one scored 80. The majority of deductions (scores 97–99) reflected incomplete distractor discussion—the system explained the governing rule and correct answer but did not affirmatively eliminate all wrong options, or minor overstatement of an analytical element. The two larger deductions (80 and 92) both reflected alternative doctrinal framing: the system selected the correct option and rejected the distractors, but grounded its analysis in a different, though defensible, legal standard than the one emphasized by the reference answer. No DescrybeLM output was flagged for wrong rule, misapplied rule, misread key fact, or contradiction. Two outputs (scoring 80 and 92) were flagged weak rule (rule statement too generic or underspecified to justify the conclusion — see Flag Definitions), reflecting the alternative doctrinal framing described above. The per-output scoring log in Appendix B provides the full detail.

Study Design

Question set

- Testing window: January–February 2026 (all runs executed within this window).
- Benchmark: The NCBE MBE Complete Practice Exam (200 multiple-choice questions; similar to a full-length MBE component of the bar exam). Source: https://store.ncbex.org/mbe-complete-practice-exam
- MBE format context (for readers): The MBE consists of 200 multiple-choice questions presented in two 3-hour sessions of 100 questions each. The exam includes 175 scored questions distributed evenly across seven subject areas (25 each): Civil Procedure, Constitutional Law, Contracts, Criminal Law and Procedure, Evidence, Real Property, and Torts, plus 25 unscored pretest questions that are indistinguishable from scored items.
- Selection rule: We purchased the NCBE MBE Complete Practice Exam and pre-committed to evaluating Questions 1–200 in order, with no exclusions and no cherry-picking.
- Reporting: The underlying questions are available for purchase from the NCBE as the MBE Complete Practice Exam. Interested readers can purchase the same set directly from the NCBE to review the underlying items and independently confirm the benchmark context and answer key (including the official correct option and explanations). We do not reproduce the question text, answer choices, or official explanations in this paper due to copyright restrictions. To generate model responses, the full question stems and answer choices were provided as inputs to each evaluated system; this paper does not assess or opine on any third-party licensing or terms-of-use implications of that processing. Instead, we report aggregated outcomes and provide per-question performance summaries without reproducing the question text, answer choices, or official answer key (including the correct letter choice). We focus on aggregated outcomes, overlaps, and failure-mode patterns, and we provide a transparency log that identifies which question numbers were missed and how outputs failed (via rubric flags), without publishing the NCBE items or answers.

Standardized instructions

Outputs were generated under a standardized “closed-book” prompt (no-external-web constraints) aligned to multiple-choice exam conditions. Each system was provided the full question stem and answer choices as input. In this paper, “no-external-web” (referred to as “closed-book” in the prompt) means: no public-web browsing and no externally enabled tools beyond the question content (as controlled via the user-facing interface settings available to evaluators). This definition does not attempt to characterize provider-side behaviors that are not observable in consumer interfaces, and it does not preclude a purpose-built system’s use of its internal, non-web primary-law corpus unless explicitly stated. Models were instructed to use only the facts provided and generally accepted legal principles; to select a single best answer

(A/B/C/D) without hedging; to avoid adding facts; and to explain the reasoning and why each other option is wrong.

Standardized prompt (verbatim)

You are answering a multiple-choice bar exam question closed-book. Rely only on generally applicable legal principles and the facts given. Do not add facts. If you must make an assumption, state it explicitly and keep it minimal. Task: Give the single best answer (A/B/C/D). Do not hedge. Explain your reasoning and explain why each other option is wrong, referencing the key legal or factual flaw.

Run conditions

- Prompt standardization: We included the standardized prompt (Appendix C) with each question for the three general-purpose LLMs. The general-purpose models required this external prompt to establish closed-book, single-best-answer, structured-explanation constraints. DescrybeLM does not operate via an external prompt; these constraints are native to the system's architecture and applied automatically through its built-in query classification and routing. DescrybeLM is designed for deep legal research by default, producing detailed multi-step outputs including tables of authorities and verified citations. When a user submits input, the system evaluates the query type. DescrybeLM detects whether each input consists of a question stem with four answer choices and if so, identifies it as an exam-style question. When the system identifies an exam-style input, it presents the user with a mode confirmation step — a standard part of DescrybeLM's query routing interface — before applying its internal exam mode constraints. This confirmation step was applied consistently across all 200 questions and introduced no per-question variation in constraints or settings. Both approaches produced equivalent operating parameters: closed-book, single best answer, structured explanation, no hedging, no external sources. We did not modify DescrybeLM's behavior per-question and did not tune settings during the run.
- Model settings: Interface defaults (i.e., where settings were not exposed to evaluators in the consumer UI, they were not observable).
- Run environment / session control: Claude Opus 4.5 and Gemini 3 Pro were run in consumer web interfaces. ChatGPT 5.2 was run via the OpenAI API Playground. All systems were accessed via paid subscription tiers. For each system, the standardized prompt was included with each question input. Each question was run as a new session/thread for the consumer UI systems and as an independent API call in the Playground. Across all four systems, this produced 800 independent evaluation runs.
- Tools / browsing (what we controlled): Web search and external tools were not intentionally enabled for any system during testing. For general-purpose models, this was controlled via the user-facing interface settings available to evaluators. Each system relied entirely on its own internal knowledge or data. DescrybeLM relies on its internal primary-law corpus by default and

does not use the public web. We cannot directly observe or verify provider-side behavior beyond what is visible in the consumer interface, and we acknowledge this as a limitation.

- DescrybeLM configuration (this benchmark): During this benchmark, DescrybeLM was evaluated using the same production configuration available to users of the platform, including its internal primary-law corpus and fine-tuned model. DescrybeLM does not use the public web or external tools by default.
- Provider-side behavior (what we cannot observe): This limitation applies beyond browsing and tools — we cannot directly observe or verify any provider-side retrieval, filtering, logging, retention, or other background behaviors beyond the user-facing interface controls and disclosures available to evaluators.
- Single pass note: Each system was run once per question. These results represent a single point-in-time observation without confidence intervals. A multi-run protocol would be needed to establish statistical reliability, and results may vary across runs due to model non-determinism and provider-side updates.

Evaluation Approach

Primary scoring rubric: A multiple-choice legal reasoning rubric aligned to bar-exam standards, emphasizing correct option selection, rule alignment, application to facts, and handling of distractors, with deductions for critical failure modes.

Blinding: Evaluation is conducted with anonymized model labels to reduce bias.

Scoring process

- Evaluator tool + blinding: A custom-built evaluator tool anonymized and randomized model outputs under blinded labels prior to scoring.
- Scoring execution: The rubric was applied consistently using a fixed grading prompt with a high-reasoning judge model (GPT-5.2 extra high reasoning). Because this judge is in the same model family as one evaluated system (ChatGPT 5.2), this is a limitation of the current evaluation, and the rubric sub-scores should be interpreted accordingly. Outputs were anonymized for scoring, the judge did not receive model identities, and the same rubric, judge prompt, and settings were applied uniformly across systems. The primary correctness findings (200/200 vs. 177–187) do not depend on the judge model. We encourage independent researchers to re-score these outputs using an alternative judge model; the rubric, prompt, and scoring methodology are fully documented to support that effort. For each question, the judge was provided (1) the reference question (stem + answer choices), (2) the official answer with explanation of why the reference answer is correct, and (3) the four anonymized model outputs. The judge then applied the rubric to score each output and assigned rubric flags.

- Disagreements / auditability: The scoring rubric was authored in advance by a human subject-matter expert and then applied uniformly by a fixed judge procedure. Because scoring was executed via a single consistent rubric application pipeline, there were no human scoring disagreements to reconcile. To address the common critique that “an LLM graded LLMs,” we emphasize (a) the rubric, and its weights/deduction triggers, were human-authored and pre-committed, (b) the judge prompt and settings were fixed and applied identically across anonymized outputs, and (c) future work may include practitioner review focused on verification burden and professional risk and additional evaluation sets beyond MBE-style multiple-choice (Future Work). We did not change the method midstream; instead we mitigate this limitation through a fixed rubric prompt, identical inputs and settings across systems, and a consistent scoring pipeline (same rubric, same judge settings, same inputs for each question).

What we measure

- Correctness: whether the system selected the correct option.
- Reasoning quality: whether the explanation identifies the governing rule and applies it correctly.
- Risk markers: We tracked high-risk failure patterns surfaced by the evaluator, including (a) dispositive fact misreads (e.g., missing "not/only/unless," timeline errors), (b) internal contradictions (e.g., the explanation argues against the chosen option), and (c) non-answers (no clear option selected). We also use the term confidently wrong to refer to cases where a system selected the wrong option while presenting a fluent, assertive explanation that does not meaningfully signal uncertainty. In practice, the confidently wrong flag was applied when an incorrect output used assertive, declarative language, stated its conclusion without qualification, and gave no observable signal that the answer might be uncertain. The flag was not applied to outputs that hedged their conclusions, acknowledged ambiguity in the facts or law, or otherwise indicated the answer could reasonably differ. Because the standardized prompt instructed all systems to select a single best answer without hedging, the flag captures a practical risk: when a system is wrong under these conditions, does the output make that error easy or hard to detect?

Results

Headline outcomes

Results summary tables:

- Table 1: Overall accuracy (% correct) and error counts across 200 questions.
- Table 2: Overall rubric score summary.
- Table 3: Overlap of incorrect answers across general-purpose models (union-of-misses set).

- Table 4: Union-of-misses questions by subject area.
- Table 5: Flag incidence among incorrect outputs (n=52).
- Table 6: Overconfidence flags across all outputs (n=200 per system)

How to read the Results: Tables 1–4 are question-level (n=200 per system). Table 5 is conditioned on incorrect outputs only (n=52 total wrong outputs across the three general-purpose models).

Table 1. Overall correctness (MBE-style benchmark; Q1–200)

System	Total questions	Correct	Incorrect	Accuracy
DescrybeLM	200	200	0	100.0%
ChatGPT 5.2	200	187	13	93.5%
Claude Opus 4.5	200	177	23	88.5%
Gemini 3 Pro	200	184	16	92.0%

Union of misses (question-level): Across the 200 questions, there were 40 unique questions where at least one general-purpose model selected the wrong option. (Because multiple models can miss the same question, this union count is not equal to the sum of errors across models.)

These results represent a directional reliability gap on this benchmark under documented conditions, not a universal performance guarantee.

Table 2. Overall rubric score summary

System	Overall Rubric Score
DescrybeLM	99.70
ChatGPT 5.2	93.41
Claude Opus 4.5	89.03
Gemini 3 Pro	91.45

Note: Scores reflect the rubric-based evaluation for the same 200 questions.

Narrative summary

Across this 200-question benchmark, the general-purpose models produced 13–23 incorrect answers each (Table 1). The misses were not confined to a single system: 40 unique questions were missed by at least one general-purpose model, and the overlap analysis shows that while many misses were unique to one model (29 questions missed by exactly one model), a meaningful subset clustered (10 questions missed by two models), with one question missed by all three (Table 3). This dispersion matters for real workflows because model disagreement does not reliably signal which output is correct; it signals that verification is required.

Rubric scores track correctness most heavily (80 points), with additional differentiation for rule alignment, application to facts, and avoidance of common distractor traps. The overall rubric-score spread (Table 2) is directionally consistent with the raw accuracy outcomes: systems that missed more questions also scored lower on rubric-based reasoning quality.

What the misses looked like

Table 5 summarizes how wrong outputs failed. Among incorrect outputs (n=52), the dominant patterns were wrong rule and misapplied rule, and most incorrect outputs were also flagged confidently wrong. In practical terms, this is the high-friction failure mode for legal users: an answer that is wrong and written in a persuasive, definitive tone increases the time and expertise required to spot the error.

We emphasize that Table 5 is conditioned on incorrect outputs only; it does not reflect the full distribution of flags across all 800 outputs. A fuller “all outputs” flag-rate report for all flag categories can be provided in future work as an aggregated summary (without item text, answer choices, or answer keys), so readers can see flag rates across all outputs, not only the incorrect subset. One finding that the incorrect-outputs conditioning obscures is worth surfacing here: the overconfidence flag appeared not only on wrong outputs but also on correct outputs from two general-purpose models. Claude Opus 4.5 received overconfidence flags on three outputs total—one incorrect (Q34) and two correct (Q5, Q95). Gemini 3 Pro received one overconfidence flag on a correct output (Q95). ChatGPT 5.2 and DescrybeLM received zero overconfidence flags across all 200 outputs. Table 6 below shows the full picture. This matters because it suggests overconfidence may be a model-level stylistic tendency in certain general-purpose systems, not merely a property of their errors. A system that applies the same assertive tone whether it is right or wrong gives practitioners less signal from output confidence alone.

Table 3. Overlap of incorrect answers across general-purpose models (union-of-misses set)

This table summarizes how many of the 200 questions were missed by exactly one, exactly two, or all three general-purpose models.

# Models Incorrect on the Same Question	# Questions
1 model	29
2 models	10
3 models	1

Unique misses (within the 40-question union set): ChatGPT 5.2 missed 7 questions that neither Claude nor Gemini missed; Claude Opus 4.5 missed 16 unique questions; Gemini 3 Pro missed 6 unique questions.

Table 4. Union-of-misses questions by subject area (40 questions)

Subject Area	# Union-Miss Questions
Contracts / Sales	9
Civil Procedure	7
Evidence	6
Real Property	6
Constitutional Law	6
Torts	4
Criminal Law & Procedure	1
Mixed	1

Flag definitions (glossary)

- Confidently wrong: Incorrect option with polished, persuasive justification that misstates law, misapplies a rule, or misreads a key fact.
- Wrong rule: Applies the wrong legal doctrine/test to the scenario.
- Misapplied rule: States a relevant rule but applies it incorrectly to the given facts.
- Misread key fact: Misreads or inverts a dispositive fact from the prompt.
- Contradiction: Internal inconsistency in reasoning that undermines the conclusion.

- Overconfidence: Uses certainty language that outstrips support under the benchmark's no-hedging constraint.
- Weak rule: Rule statement is too generic/underspecified to justify the conclusion.

Table 5. Flag incidence among incorrect outputs (n=52)

Important: This table is conditioned on outputs that were already wrong (n=52). Flag rates describe how wrong outputs failed, not how often each model produced each flag across all 200 questions.

Flag

- ChatGPT 5.2 (n=13 wrong outputs)
- Claude Opus 4.5 (n=23)
- Gemini 3 Pro (n=16)
- Total (n=52)

Flag	**ChatGPT 5.2**	**Claude Opus 4.5**	**Gemini 3 Pro**	**Total**
Confidently wrong	12	21	16	49
Wrong rule	10	19	13	42
Misapplied rule	7	16	9	32
Misread key fact	1	0	1	2
Contradiction	0	0	1	1
Overconfidence	0	1	0	1

Two DescrybeLM outputs (Q79, Q188) were flagged weak rule on correct answers; these are reported in the DescrybeLM score distribution discussion above and are not included in this table.

See Appendix B for the per-output log of incorrect answers and associated flags.

Table 6. Overconfidence flags across all outputs (n=200 per system)

System	Total Wrong Outputs	Overconfidence on Wrong Outputs	Overconfidence on Correct Outputs	Total Overconfidence Flags (all outputs)
DescrybeLM	0	0	0	0
ChatGPT 5.2	13	0	0	0
Claude Opus 4.5	23	1 (Q34)	2 (Q5, Q95)	3
Gemini 3 Pro	16	0	1 (Q95)	1

Note: Overconfidence flags on correct outputs were not surfaced in Table 5 because that table is conditioned on incorrect outputs only. The full scoring log was reviewed across all 800 outputs to produce this table. Claude Opus 4.5 and Gemini 3 Pro share a flag on Q95, suggesting a common response tendency on that question.

Failure Modes

The failure-mode numbers from Table 5 are summarized above. This section describes what those patterns look like in practice, generalizing the recurring error types observed across the 52 incorrect outputs. These descriptions are not tied to individual questions but reflect the categories of reasoning errors that appeared across models and subject areas.

Wrong rule patterns

The most common failure was applying an entirely wrong legal doctrine to the facts presented. These errors took several recurring forms:

Selecting a weaker legal defense or theory when a stronger, dispositive one was available. In negligence questions, models chose "no breach" (the defendant wasn't careless) when the stronger and correct defense was "no duty" (the defendant owed no obligation to the plaintiff at all). This pattern appeared across tort questions where a threshold element would have resolved the case before reaching the element the model chose to analyze.

Applying the wrong source of law in a federal/state choice-of-law context. Models applied state evidentiary rules in federal proceedings where a specific federal rule governed, or vice versa. In questions testing which body of law controls privilege waiver or procedural defenses in federal diversity cases, models defaulted to the wrong sovereign's rules, producing analyses that were internally coherent but built on the wrong legal foundation.

Invoking a doctrine that does not apply to the procedural or factual posture presented. Models applied political question doctrine to challenges involving the constitutionality of a federal statute, where the issue was a standard separation-of-powers question with judicially manageable standards. Models applied subject-matter waiver to inadvertent disclosures where the doctrine is limited to intentional disclosures. Models applied gap-filler pricing doctrines to contracts where the price term was already established through the parties' conduct.

Fabricating or overstating distinctions between related doctrines. In questions testing whether two parallel doctrines (such as evidentiary and constitutional protections) share the same requirements, models invented distinctions that do not exist in the law, concluding that one protection applied while the other did not, when in fact both require identical showings.

Mischaracterizing a legal status or classification. In premises liability questions, models classified an entrant as a trespasser based on a posted sign, even though an employee's express direction to enter that area superseded the general restriction. In comparative fault questions, models compared the plaintiff's fault to the aggregate fault of all defendants rather than to each defendant individually, as the governing rule required.

Applying the wrong remedy or measure of damages. Models awarded reliance damages when expectation damages were appropriate, or limited recovery to nominal damages when the plaintiff had suffered a measurable loss. In contract formation questions, models confused the mechanism that made an obligation enforceable (such as detrimental reliance creating an irrevocable offer) with the remedy available once the obligation was breached.

Misapplied rule patterns

The second most common failure involved identifying a relevant legal doctrine but applying it incorrectly to the facts:

Using the wrong starting point for a time-based legal test. In statute of frauds questions, models measured the one-year period from when the offer was communicated rather than when the contract was formed (upon acceptance), producing an incorrect conclusion that the contract fell within the statute.

Reversing who bears a burden or which party a rule protects. Models correctly identified the governing standard but applied it in the wrong direction, concluding that a protection applied to shield a party when the rule's requirements were not met, or that a bar applied when the triggering condition was absent.

Treating a failed procedural safeguard as converting one legal category into another. When a party failed to meet the requirements for an exception (such as the promptness requirement for protecting an inadvertent disclosure), models concluded that the failure transformed the disclosure into an intentional one, triggering broader consequences that only attach to truly intentional conduct.

Misapplying the scope of an invitation or permission. Models correctly identified the relevant premises liability framework but drew the boundary of permission too narrowly or too broadly, ignoring facts that expanded or limited the entrant's authorized area.

Misread key fact patterns

A smaller but operationally significant set of errors involved misreading or overlooking a dispositive fact in the question:

Overlooking a fact that established a key legal element. In contract questions, models missed that a party had been told specific information (such as a posted rate or a website reference) that established notice or assent, then treated a contract term as missing when it was actually present in the facts.

Failing to recognize a timing fact that resolved a procedural question. Models overlooked that a specific time interval fell within an applicable deadline, concluding that a right had been waived when it had in fact been timely preserved.

Contradiction patterns

In rare cases, a model's own reasoning contradicted its conclusion:

Reasoning toward the correct answer, then selecting the wrong one. In at least one instance, a model walked through the analysis, reached the correct legal conclusion in its reasoning, but then selected a different answer choice, creating an internal inconsistency visible in the output itself.

General-purpose model patterns

It is worth noting that these error patterns appeared exclusively in the general-purpose model outputs. DescrybeLM was not flagged for wrong rule, misapplied rule, misread key fact, or contradiction on any of its 200 outputs. While this benchmark cannot establish why the general-purpose models exhibited these patterns, the consistency of the failure modes across three independently developed systems suggests they may reflect structural tendencies of general-purpose LLMs when applied to legal reasoning tasks on this benchmark, rather than idiosyncratic errors specific to any one model. Purpose-built systems with access to curated legal data and domain-specific reasoning pathways may be less susceptible to these patterns, though further testing across broader task types would be needed to confirm that hypothesis.

Across nearly all of these error types, the outputs exhibited the same surface-level quality: fluent prose, structured analysis, and definitive tone. The wrong rule outputs in particular were dangerous precisely because they read like competent legal reasoning. A model that applies the wrong legal standard but does so cleanly and confidently produces an output that requires the reader to independently verify which standard governs before the error becomes visible. This is the core of the verification burden problem: the errors do not announce themselves.

Implications

This benchmark is a controlled multiple-choice test, but the reliability patterns are directly relevant to first-pass analysis workflows where users must decide what to trust and what to re-check.

Practical takeaways

1. On this benchmark, the purpose-built system showed lower error rates and higher rubric-scored explanation quality under the documented constraints. DescrybeLM answered all 200 questions correctly, compared to 177–187 for the general-purpose models. For exam-style legal reasoning tasks, and for workflows where incorrect first-pass answers create downstream risk, a purpose-built legal system showed stronger performance on this benchmark. Rubric-scored reasoning quality followed the same pattern: DescrybeLM scored 99.70% compared to 89.03%–93.41% for the general-purpose models, with deductions driven primarily by incomplete distractor discussion and alternative doctrinal framing rather than rule errors or fact misreads. The prompt was designed to give the general-purpose models the clearest possible instructions for this task — conditions intended to support accurate, well-reasoned responses. Those same parameters are native defaults in DescrybeLM, meaning the prompt was helping the general-purpose models operate under equivalent conditions, not constraining them.
2. Treat “sounds right” as a risk factor, not a comfort. The most operationally costly failures are outputs that are wrong while sounding definitive (confidently wrong). They increase verification burden because they do not invite the user to slow down.
3. Overconfidence may be a model-level tendency, not just an error pattern. A review of the full scoring log identified overconfidence flags on correct outputs from two general-purpose models — Claude Opus 4.5 and Gemini 3 Pro — that Table 5 did not surface because it is conditioned on incorrect outputs only. ChatGPT 5.2 and DescrybeLM received zero overconfidence flags across all 200 outputs. This matters because a system that uses the same assertive tone whether it is right or wrong provides less signal to practitioners. A confident tone on one correct output does not indicate the next output will also be correct. DescrybeLM’s clean overconfidence record across all 200 outputs is consistent with its broader reliability profile on this benchmark (see Table 6).
4. Model disagreement does not solve verification. The 40-question union-of-misses and overlap results show that misses are distributed across models. Disagreement is a signal that the problem is brittle, not a reliable tiebreaker.
5. Wrong rule and misapplied rule failures dominate. Among incorrect outputs, wrong rule and misapplied rule were the most common patterns. In practice, these are exactly the errors that can “look professional” and survive superficial review.

6. In legal settings, evaluation should weight reliability signals and auditability at least as heavily as writing quality. The value is not eloquence—it's defensible reasoning, predictable failure modes, and verification-friendly outputs.
7. Whether higher first-pass accuracy on multiple-choice benchmarks translates to reduced verification effort in practice is untested in this study and remains an open question.

How to evaluate legal AI tools

When evaluating legal AI systems, we recommend assessing four dimensions: (1) run-condition reproducibility, (2) outcome quality beyond correct answers, (3) verification burden, and (4) failure-mode transparency. These dimensions generalize the reliability lens introduced earlier in this paper into a practitioner-facing evaluation framework. The checklist in Appendix D provides a detailed, repeatable version.

Minimum disclosure standard

If a vendor publishes an evaluation, it should, at minimum, disclose: the dataset/source, selection rule, versions tested, run conditions (UI/API + defaults), whether tools/browsing were enabled, the scoring rubric, and the blinding/scoring process.

Limitations and Future Work

This evaluation is intentionally narrow and controlled.

Limitations

- Benchmark format: MBE-style multiple-choice questions are a constrained test format. They do not fully represent end-to-end legal work (research, drafting, jurisdiction-specific analysis, factual development, strategy).
- No-external-web conditions: This benchmark evaluates responses under standardized no-external-web instructions. It does not measure tool-assisted research performance, citation accuracy, or authority-checking behavior.
- Copyright constraints: Because NCBE items cannot be reproduced, this paper reports aggregated outcomes, overlap patterns, and failure-mode logs keyed to question number + subject rather than publishing item text or answer keys.
- Evaluator observability: Provider-side behavior for general-purpose models is not directly observable (see Run Conditions).

- Single-pass design: Each system was run once per question. Because LLM outputs are non-deterministic, results may vary across runs, and this study does not report confidence intervals or variance data. A multi-run protocol is planned for future evaluations.

Future work

- Research-grounded evaluations: Expand to tasks where traceability and authority checking are central (e.g., identifying controlling authority, distinguishing binding vs persuasive, checking citations).
- Practitioner review: Practitioner feedback may be incorporated in future iterations to assess verification burden and professional risk in real workflows.
- Regression testing over time: Repeat the benchmark periodically to measure drift as models and products change.
- Open-book evaluation: This benchmark was conducted under closed-book, no-external-web conditions. Whether enabling web browsing or external tool access would improve general-purpose model accuracy is an open question. Web access could provide additional legal reference material, but it could also introduce noise, hallucinated citations, or reliance on secondary sources. A future evaluation comparing open-book and closed-book performance across all systems is planned.
- Beyond MBE-style: Add additional controlled question sets and real-world scenarios beyond MBE-style multiple-choice (e.g., research-grounded tasks, jurisdiction-sensitive questions, and workflow-based evaluations), with clear selection rules and disclosure. Going further, invite legal academics to submit their model exams as benchmark tasks to evaluate which tools perform best. Contact info@descrybe.com to discuss opportunities.

Conclusion

This paper evaluates a narrow question under controlled constraints: how reliably do different systems select the correct answer and explain it using defensible legal reasoning on an MBE-style benchmark? On this 200-question set, DescrybeLM achieved 200/200 correct answers and the highest rubric score, while general-purpose models missed 13–23 questions each and exhibited failure modes dominated by wrong rule and misapplied rule reasoning, often delivered in a confidently wrong style.

The broader takeaway is not that general-purpose models are unusable, but that legal work punishes confidently wrong: the cost is not just the error itself, but the verification burden and the downstream risk created by persuasive-but-incorrect outputs. For buyers and practitioners, the practical path forward is to evaluate legal AI systems on reliability, auditability, and verification burden, not on fluency alone, and to favor architectures and workflows that make it easier to detect and correct errors before they propagate.

We invite independent replication: the question set, standardized prompt, and scoring rubric structure are fully specified in this paper, and the NCBE source materials are commercially available. Due to model non-determinism and provider-side updates, exact numerical replication is unlikely; directional replication (rank ordering of systems, approximate accuracy ranges) is the appropriate expectation for independent testers. Readers who wish to evaluate reasoning quality independently may apply their own rubric to the same outputs; accuracy results (correct/incorrect) are verifiable using the commercially available answer key.

Appendices

Appendix A: Evaluation Framework (Rubrics)

This v1 white paper uses the bar-exam multiple-choice rubric below. (Other rubric modes exist for separate, non-MBE evaluations, but are not used in this paper.)

A1. Bar-exam multiple-choice rubric (MBE-style; exam-aligned)

Our scoring rubric is a bar-exam aligned multiple-choice scoring rubric designed to prioritize correct option selection while still distinguishing between (a) answers that generalize reliably and (b) answers that are correct for shaky reasons.

Total: 100 points

Option Correctness (0–80)

- 78–80: Correct option clearly selected; no flip-flop.
- 70–77: Correct option selected; reasoning thin/imprecise but broadly baseline-consistent.
- 55–69: Correct option selected but with a material confusion (wrong standard/doctrine) that could plausibly lead to wrong answers in similar questions.
- 0–54: Wrong option selected or no clear option.

Rule Alignment / Accuracy (0–10)

- 9–10: States the core governing rule consistent with the reference baseline.
- 6–8: Mostly consistent; minor imprecision that wouldn't change outcomes.
- 0–5: Material rule conflict with baseline.

Application to Facts (0–7)

- 6–7: Connects key facts to the rule in a baseline-consistent way.
- 3–5: Minimal or somewhat conclusory application, but not wrong.
- 0–2: Misapplies rule to facts or relies on invented/misread facts.

Distractor Handling & Trap Avoidance (0–3 bonus)

- 3: Correctly knocks out at least one tempting distractor with a rule-based reason.
- 1–2: Mentions distractors but shallow.
- 0: No distractor discussion (no penalty) or distractor discussion contradicts the correct answer.

Global deductions (serious problems only):

- Hallucinated key fact (inventing a dispositive fact): –10 to –40.

- Dispositive fact misread (e.g., misses “not/only/unless,” wrong timeline): –10 to –30.
- Contradiction (explanation argues against the chosen option): –10 to –25.
- Non-answer (no clear option): Option Correctness cannot exceed 20.

Scoring principle: Do not “death-by-nuance” penalize. If the candidate reaches the correct option with generally sound baseline reasoning, do not over-penalize stylistic differences.

Appendix B: Per-question performance summary (transparency log; no NCBE content)

To support transparency while respecting NCBE copyright, the accompanying supplemental file (DescrybeLM_Appendix_B_Scoring_Log.xlsx) contains a per-output scoring log covering all 800 model outputs (one row per system per question) reporting rubric score, reasoning alignment, reliability classification, and flag incidence. These logs include question number, system, correct/incorrect, rubric score, and flag incidence, without reproducing NCBE question text, answer choices, or the official answer key. Incorrect outputs are highlighted in yellow; DescrybeLM outputs with rubric scores below 100 are highlighted in blue.

Appendix C: Standardized prompt used for general-purpose models

The following standardized prompt was used to generate responses from the general-purpose LLMs in this evaluation.

You are answering a multiple-choice bar exam question closed-book. Rely only on generally applicable legal principles and the facts given. Do not add facts. If you must make an assumption, state it explicitly and keep it minimal. Task: Give the single best answer (A/B/C/D). Do not hedge. Explain your reasoning and explain why each other option is wrong, referencing the key legal or factual flaw.

(Note: The paper does not reproduce NCBE items; this appendix reproduces only the evaluator instruction prompt.)

Appendix D: Evaluation checklist (verification-first)

This checklist provides a repeatable framework for evaluating legal AI systems in a way that aligns with professional risk.

Run conditions & reproducibility

- Are the tested model versions, interfaces (UI vs API), and default settings documented?
- Were browsing/tool features intentionally enabled or disabled (and disclosed either way)?

- Is there a pre-commitment selection rule (e.g., “Q1–200 in order”) that prevents cherry-picking?

Outcome quality (not just correct answers)

- Does the system select the correct outcome under defined constraints?
- Does it apply the right governing rule and connect it to the key facts?
- Does it avoid internal contradictions and “best-guess” fill-ins when uncertain?

Verification burden

- Can a reviewer quickly see what the answer depends on?
- When the system is wrong, is it obviously wrong, or does it sound definitive?
- Are there guardrails that reduce confidently wrong outputs (e.g., strong uncertainty signaling, traceability, structured reasoning steps, provenance)?

Failure-mode monitoring

- Does the vendor report failure modes (e.g., wrong rule, misapplied rule) and how frequently they appear?
- Are examples representative (selection rule disclosed) and supported by an audit log?

Fit for your workflow

- For legal aid / high-volume teams: does the tool reduce time-to-verify, not just time-to-draft?
- For firms: does it support review workflows (team QC, escalation, audit trails)?
- For integrators: can you measure reliability over time (batch tests, regression testing, change logs)?