Using Differences in State Law to Test Whether LLMs Reason or Remember

Colin Doyle [*1] **Aaron D. Tucker** [*2]

Abstract

When a large language model correctly applies a legal rule to facts, it's hard to tell whether the model has reasoned its way to the answer or has followed a context-specific pattern learned from its training data. To disentangle these two accounts, we exploit state-by-state legal variation in the United States, treating a legal rule's prevalence across states as a proxy for its prevalence in LLMs' training data. Legal rules can be placed along a spectrum: *majority rules* adopted by most states, *minority rules* adopted by few states, and *hypothetical rules* adopted by none. Drawing upon the legal topics of comparative fault, non-compete agreements, and hearsay, we evaluate LLMs' performance applying a majority rule, a minority rule, and a hypothetical rule of similar complexity. If a model's process of applying law to facts approximates legal reasoning, its accuracy should be consistent across all three rule conditions. But if a model relies upon context-specific patterns internalized during training, the model's performance should vary based on the prevalence of the rule within its training data. In two of three legal domains, we find that models perform best on majority rules, worst on minority rules, and at an intermediate level on hypothetical rules. In the remaining legal domain, models consistently and accurately apply all three rules. These results suggest that LLM performance at legal reasoning tasks is shaped, in part, by the prevalence of similar legal rules in models' training data. Future research can adopt this method of using state-by-state variation in legal rules as a principled way to distinguish legal reasoning from context-specific pattern matching.

[*]Equal contribution [1]Loyola Law School, Los Angeles, California, USA [2]FAR.AI, Berkeley, California, USA. Correspondence to: Colin Doyle <colin.doyle@lls.edu>.

Oral Presentation + Outstanding Paper Award at Proceedings of the AI for Law Workshop at the 43rd International Conference on Machine Learning, Seoul, South Korea, 2026. PMLR 356, 2026. *Copyright 2026 by the author(s).*

1. Introduction

Frontier models have demonstrated strong performance on some legal tasks, prompting efforts to deploy models for consequential, complex legal work (Magesh et al., 2025; Doyle, 2025). Leading legal benchmarks — like LegalBench (Guha et al., 2023) and the legal tasks contained within GDPval (Patwardhan et al., 2025) — appear to validate models' capacity to do legal reasoning work. But other research indicates that models' legal reasoning capabilities may be more fragile and erratic than the benchmarks indicate (Doyle & Tucker, 2025; Blair-Stanek et al., 2024; 2023).

The core problem is familiar to the NLP community but under-examined in legal AI evaluations: When a model gives the right answer to a legal question, it is impossible to tell whether the model reasoned its way to a conclusion or simply reproduced a context-specific pattern memorized in training. Under current legal benchmarks, either process would produce the same result but would represent dramatically different measures of a generalized legal reasoning capability. A model that produces correct answers to legal questions because it has internalized a pattern that approximates legal reasoning can be deployed to new legal contexts. But a model that produces correct answers to legal questions by following context-specific patterns may fail spectacularly when deployed in contexts that differ from its training data.

Our experiment is rooted in a straightforward premise: If a model's accuracy in applying a legal rule to facts depends upon the prevalence of that legal rule within its training data, then the model's accuracy should degrade when the model is prompted to apply legal rules that are less prevalent within the training data. And vice versa, if a model has internalized a generalizable method of legal reasoning, the accuracy of the model's answers should not depend on the prevalence of the legal rule within its training data.

To perform this experiment, we exploit state-by-state legal variation in the United States as a proxy for training-data prevalence, constructing three sets of legal reasoning tasks drawn from legal domains in which states have adopted different rules for resolving the same questions. For each run, we present large language models with a factual scenario and one of three versions of a legal rule: a *majority rule* adopted by the greatest number of jurisdictions; a *minority*

rule adopted by the smallest number of jurisdictions; and a *hypothetical rule* that is plausible and coherent as a legal standard but is not the actual rule of any jurisdiction.

We find that models perform best on majority rules, worst on minority rules, and at an intermediate level on hypothetical rules, providing support for the conclusion that a frontier model's accuracy with applying a legal rule can depend upon the rule's prevalence within the model's training data. Our method – using cross-state variation as a proxy for a rule's representation in the models' training data – can be adopted by future research for evaluating LLMs' legal reasoning capabilities more broadly. This method is practical, designed to resist training-data contamination, and grounded in established practices of assessing humans' legal-reasoning competence. To help future research, we have publicly released our experiment's dataset as the RuleSwap dataset on Huggingface: https://huggingface.co/datasets/aarondtucker/ruleswap.

2. Background

2.1. Shortcuts and Spurious Correlations

Machine-learned models like LLMs use data and large amounts of computation to find programs that perform well according to a specified metric of success. Across a variety of tasks and domains, machine-learned models find ways to attain high performance by exploiting correlations in data – sometimes instead of learning an intended capability. This can result in programs that perform well on a particular metric but whose shortcuts limit their robustness or generalization (Geirhos et al., 2020). For example, computer vision models trained on ImageNet often based their predictions on textures in the image that correlated with the correct class, rather than on features like shape or relative location (Geirhos et al., 2018). The language model BERT performed well when asked to infer if one sentence entailed another, but the model failed when sentences lacked simple heuristics like having overlapping subsequences (McCoy et al., 2019). And newer multimodal reasoning models have performed well on vision-based radiological tests – without the models ever seeing the test's x-rays (Asadi et al., 2026).

Our experiment tests a closely related issue: whether a language model's accuracy in applying a legal rule depends upon how frequently the rule appears within its training corpus. If accuracy depends upon rule prevalence, the model has not learned to apply legal rules in a fully generalized way. The legal skills that LLMs exhibit are hard to pin down. They can be understood as a difficult-to-separate blend of the results of both purposeful training and the emergent capabilities of a complex system (Wei et al., 2022). For instance, legal reasoning abilities may be based in generalization from instruction-following training (Ouyang et al., 2022), implicitly learning legal concepts from pretraining and supervised finetuning (Brown et al., 2020), and more general reasoning-like abilities (Lambert et al., 2025; Guo et al., 2025; Kojima et al., 2022). Nonetheless, shortcuts and spurious correlations may affect an LLM's outputs. For example, in next-token prediction the model could learn a pattern for saying that "X is an enforceable contract" or "X is not an enforceable contract" based on incidental features of commonly represented text rather than the specific definitions of a legal rule provided in the prompt. This would result in brittle reasoning performance, since applying the legal concept of contract enforceability would depend on context-specific language patterns rather than the rule's actual content. In contrast, a model with genuine legal reasoning ability would reliably apply legal rules without relying on incidental features of the text.

2.2. LLMs' Legal Reasoning Capabilities

The empirical "Law & AI" literature is only beginning to answer fundamental questions about LLMs' legal reasoning capabilities. It remains unclear how much LLMs have developed an emergent legal reasoning ability, how much their success depends on the presence of similar examples in their training data, and how much strong performance on one type of legal task can generalize to other legal tasks. A key finding from the broader literature on LLM performance is that it is not uniform across tasks: LLM capabilities are a "jagged frontier" — they excel at some tasks but fail at others of comparable or even lesser complexity (Dell'Acqua et al., 2026).

Empirical research into LLMs' performance on legal reasoning tasks finds this same jagged pattern. Some research has found that LLMs perform nearly as well as, or even better than, law students and practicing lawyers on certain legal tasks (Choi & Schwarcz, 2025; Nay et al., 2024). Other studies have revealed how LLMs fall short. LLMs struggle with routine text-based tasks common in legal practice and make consistent errors in basic statutory reasoning tasks (Blair-Stanek et al., 2023; 2024). These statutory reasoning errors can become even more pronounced when the model is given hypothetical statutes that could not have appeared in its training data — suggesting that the model's performance was dependent, in part, on context-specific pattern matching rather than an approximation of legal reasoning (Blair-Stanek et al., 2023). In one set of experiments testing LLMs' structured legal reasoning ability across different state-level legal domains, researchers gave the models relevant instructions from legal practice guides and then asked them straightforward legal questions (Doyle & Tucker, 2025). LLMs frequently failed to answer correctly even when the practice guide provided clear instructions for applying the law. The errors were not confined to areas of legal ambiguity or doctrinal complexity but were inconsistent and unpredictable,

with more advanced "reasoning" models performing worse than smaller language models in some contexts.

3. Methods

3.1. Prompting

This experiment evaluated LLMs' performance on legal reasoning questions drawn from three areas of law: comparative fault in tort law, hearsay exceptions in the law of evidence, and the enforceability of noncompete agreements in employment contracts.

Each LLM query was constructed from three components: a prompt, a rule, and a question. The prompt, held constant across all trials, instructed the model that it was acting as a legal expert asked to answer a legal reasoning exam question, and directed it to conclude its response with either "YES" or "NO." The model was told it was answering an exam question for two reasons: to reduce the chance that guardrails against giving legal advice would suppress responses and to give the model a plausible reason to apply an unfamiliar or fictional rule rather than treat the prompt's premise as an error to correct (Doyle, 2024). On a law school exam, a test-taker would expect to be asked to apply a hypothetical law to a fact pattern. Without this context, a model may resist applying a fake law to facts — not because the model is incapable of doing so but because the model is instead looking to correct the prompt's mistaken understanding of what the law is.

The rule component consisted of the text of statutes and caselaw governing a particular legal rule. In each of the three areas of law tested in this experiment, states have adopted different rules to address the same legal questions. For each area of law we selected and edited majority, minority, and hypothetical versions of the governing rule. Rules were drafted to resemble authentic legal materials. Statutory text was copied or modeled upon real state codes and the Federal Rules of Evidence. Caselaw excerpts were copied or adapted from real legal opinions.

Because significant variations in the length or complexity of the different rules could confound the results, we selected and edited the three different rules within each area of law to maintain similar word counts and levels of complexity. For example, California has adopted a minority rule that noncompete agreements are unenforceable, but we did not select California's rule for the experiment because that rule is far simpler than other states' noncompete agreement rules that include multiple factors to consider and detailed lists of exceptions.

Because providers do not disclose training data, we cannot directly measure a legal rule's frequency in any model's training corpus. State-adoption count is therefore a proxy for training-data prevalence, resting on the assumption that the more prevalent rule appears more frequently in caselaw, online study materials, and the secondary literature. This assumption would not hold for idiosyncratic minority rules that attract outsized scholarly attention and practitioner commentary precisely because they are unusual. To mitigate this risk, we avoided selecting minority rules that have received disproportionate attention.

The question component consisted of a fact pattern and a binary legal question. For each of the three areas of law, we wrote between 8 and 11 fact patterns (8 for damages, 10 for hearsay, 11 for noncompete), each designed to produce different legal results under the different rules. Models were selected from three commercial providers: three Anthropic models (Claude Haiku 4.5, Claude Sonnet 4.6, and Claude Opus 4.7), three OpenAI models (GPT-5.4 Mini, GPT-5.4, and GPT-5.5), and three Google models (Gemini 3.1 Flash Lite, Gemini 3.5 Flash, and Gemini 3.1 Pro). Within each provider, three models were selected to span a range of capabilities from smaller, faster models to "reasoning" models that deliberately use "thinking" tokens to work through an answer before responding. Each query-model combination was run 10 times. The prompts were provided to the models via the API, with the prompt provided as the user input, with no system prompt. Extended reasoning was enabled for the models that support it. Because some providers do not allow temperature parameters to be adjusted for certain models, all models were queried at temperature of 1.0 to keep settings consistent across models. Provider-specific settings are detailed in Appendix Table 1. Our dataset is available on Huggingface: https://huggingface.co/datasets/aarondtucker/ruleswap.

3.2. Legal Domains and Rules

3.2.1. Damages

When one person negligently injures another, tort law must provide an answer to two questions: can the injured party (the plaintiff) recover compensation from the other party (the defendant), and if so, how much? Complications arise when multiple defendants are at fault or when the plaintiff is also partly to blame for their injuries. Over time, jurisdictions have developed different rules. Traditionally, if a plaintiff was even a little bit negligent, they would be barred from recovery, with some exceptions. Today, almost all states have relaxed this rule by adopting some form of *comparative fault*, allowing plaintiffs to recover even if they are partly at fault but reducing their damages in proportion to their share of fault. Most states still bar recovery entirely once a plaintiff's fault reaches or crosses a 50% threshold. These rules interact with other doctrines: *assumption of risk*, a common-law defense that bars recovery when a plaintiff knowingly accepts a risk, and the *last clear chance* rule,

which allows a negligent plaintiff to recover despite their negligence if the defendant had a final opportunity to prevent the harm and failed to take it. The majority, minority, and hypothetical rules each address the issues of comparative fault, allocation of damages, assumption of risk, and last clear chance. The factual scenarios in this dataset describe a jury's factual findings — including each party's percentage of fault — and ask the model to apply the governing rule to determine whether the plaintiff can recover.

Majority Rule. The majority rule presents a modified comparative fault statute modeled on Pennsylvania law, 42 Pa. Stat. and Cons. Stat. Ann. § 7102, under which a plaintiff may recover so long as their negligence does not exceed the defendant's — in effect, a 50% fault threshold. The doctrines of "assumption of risk" and the "last clear chance" rule are no longer in effect.

Minority Rule. For the minority rule, we selected the "slight-gross" comparative fault standard from South Dakota, S.D. Codified Laws § 20-9-2, the only state to currently have this rule in place. Under this rule, a plaintiff is barred from recovery if their negligence is more than "slight" in comparison to the defendant's negligence. This is a much less plaintiff-friendly rule than the majority rule. South Dakota also retains the doctrines of assumption of risk and last clear chance.

Hypothetical Rule. The hypothetical rule borrows most of its text from the majority rule and is a modified comparative fault regime. Under the majority rule, a plaintiff is barred from recovery if they are more than 50% at fault. In tort cases with multiple defendants, the plaintiff's fault is compared with the defendants' collective fault. So long as the plaintiff isn't more at fault than the defendants as a group, the plaintiff can recover. The hypothetical rule rewrites the test to be that a plaintiff is barred from recovery if no single defendant is more at fault than the plaintiff. In a case with multiple defendants, the plaintiff's fault must be compared with each defendant's fault, one at a time. So long as at least one defendant is more at fault than the plaintiff, the plaintiff can recover. This rule is plausible and logical but has not been adopted by any state. Assumption of risk is a full defense. The last clear chance rule remains in force in its historical "British" form, although England has abandoned the rule and no U.S. state currently follows it. Under this plaintiff-friendly variant, a defendant can be held liable if they would have had the last clear chance to avoid an accident, but they stripped themselves of that chance through their own prior negligence.

The hypothetical rule was drafted to conform to the language of the majority rule as much as possible for two reasons. First, it removes a potential confound: differences in how the two rules are phrased. Second, because the text of the hypothetical rule stays as close to the majority rule as possible while producing a different correct answer, a model that reaches the right answer offers stronger evidence that it was not swayed by surface similarity to training data but performed genuine legal reasoning.

Fact Patterns. The comparative fault fact patterns consist of eight variations on a car accident scenario in which two or three cars have crashed into one another. Across the scenarios, a jury has returned a verdict assigning a percentage fault to each of the drivers. Each scenario holds the underlying facts as constant as possible while varying the allocation of fault. The scenarios include a range of fault distributions that can produce divergent outcomes across the three rules: a plaintiff bearing minimal fault (5%), a plaintiff bearing fault just below the majority rule's 50% bar (45%), a plaintiff at exactly 50%, and a plaintiff bearing more than half the fault (55%). Two scenarios introduce a second defendant, testing how each rule handles multi-party fault allocation — a dimension on which the majority and hypothetical rules diverge most sharply, since the hypothetical rule compares the plaintiff's fault against each defendant individually rather than collectively. The final two scenarios present last-clear-chance situations, one in which the doctrine operates under standard conditions and one structured to trigger only the hypothetical rule's distinctive rule under which a defendant who forfeits the last clear chance through prior negligence remains liable.

3.2.2. Noncompete Agreements

When an employee leaves a job, their former employer may wish to prevent them from immediately working for a competitor. A noncompete agreement is a contractual clause in which the employee promises not to compete with the employer, often for a defined period of time within a particular industry and geographic area. Courts asked to enforce these agreements must balance competing legal values. The employer has a legitimate interest in protecting trade secrets, client relationships, and competitive advantages developed at the employer's expense. But the employee has a right to earn a living. And there's a broad public-policy interest in facilitating competitive labor markets. Because these agreements function as restraints on commerce, courts do not enforce them automatically as written. Most jurisdictions require that the restrictions be reasonable, and many jurisdictions impose additional requirements. When noncompetes are found to be unreasonable, jurisdictions diverge in how a court is permitted to address the issue. Some courts are allowed to rewrite the terms of the contract to be reasonable; other courts are permitted only to strike unreasonable provisions while enforcing the rest (referred to as "blue-lining"); and other courts are not allowed to modify the

agreement but must void the agreement in its entirety. The majority, minority, and hypothetical rules differ both in what kinds of noncompete agreements are permissible and in how a court is permitted to address unreasonable noncompete agreements. The fact patterns present different employment situations and noncompete agreements with different terms. In each scenario, the model must determine either whether the noncompete is enforceable under the governing rule or whether the court is permitted to rewrite the terms of the contract in a particular way.

Majority Rule. The majority rule is adapted from Michigan's statute and caselaw. Mich. Comp. Laws § 445.774a. Noncompete agreements are enforceable if reasonable in duration and geographic scope, but are only enforced when an employee would be gaining some kind of unfair advantage, like poaching clients or providing a new employer with confidential information. Courts are authorized to rewrite the terms of unreasonable noncompete agreements to make them reasonable.

Minority Rule. The minority rule is taken from Wyoming. Under state statute, noncompete agreements entered into after July 1, 2025 are void, with narrow statutory exceptions for trade secret protection and executive personnel. Wyo. Stat. Ann. § 1-23-108. The rule explicitly prohibits courts from rewriting the terms of an unreasonable contract or removing ("blue-lining") unreasonable terms from an otherwise valid contract. Instead, if any of the terms of the contract are unreasonable, the entire contract is invalid and unenforceable.

Hypothetical Rule. The hypothetical rule, modeled loosely on proposed legislation in Virginia, contains a mix of rare and novel elements. 2026 Virginia Senate Bill No. 170, Virginia 2026 Regular Session. Noncompete agreements are valid only if they include a severance-pay provision that provides continued compensation to the employee for the duration of the restricted period. There are exemptions for restaurant workers and lower-wage earners, including student interns. The court cannot rewrite unreasonable terms of a contract, but the court can remove unreasonable terms from the contract.

Fact Patterns. The noncompete fact patterns consist of eleven scenarios built on an employment dispute between an employee looking to change jobs and their employer seeking to enforce a noncompete agreement the employee signed. The core agreement used across the scenarios is a realistic, multi-provision noncompete that includes a one-year restriction, a severance-pay provision, and standard confidentiality provisions. Individual scenarios modify one variable at a time: whether the agreement was signed at the outset of employment or mid-employment without additional consideration; whether the restricted period is one year (reasonable) or ten years (unreasonable); whether the agreement includes the severance-pay provision required by the hypothetical rule; and whether the employee is a salaried professional or a low-wage intern. Two scenarios test the treatment of restaurant workers — a category exempted under the hypothetical rule. And two scenarios test whether courts may rewrite or sever a term of the agreement rather than voiding the agreement entirely.

3.2.3. HEARSAY

At trial, witnesses testify from personal, first-hand knowledge. The hearsay rule addresses what happens when a party wants to introduce a statement that was made outside of court. Out-of-court statements are presumptively inadmissible as hearsay because the person who made the statement is not on the witness stand, cannot be cross-examined, and was not under oath when they spoke. The rationale is that out-of-court statements are vulnerable to problems with memory, perception, and sincerity (Imwinkelried, 2016). But the rules of evidence also recognize exceptions for certain kinds of out-of-court statements in which the ordinary risks of hearsay are minimized. These exceptions allow otherwise-inadmissible statements to be admitted into evidence. This experiment focuses in particular on the "present-sense impression" and "excited utterance" exceptions. The "present-sense impression" exception allows a statement made while a person is actively perceiving an event, or immediately after, to be admitted into evidence, and the "excited utterance" exception allows a statement made while a person is "under the stress of excitement" caused by an event to be admitted into evidence (Imwinkelried, 2009). The majority, minority, and hypothetical rules differ in whether they recognize these exceptions and, if they do recognize an exception, the conditions under which the exception applies. The fact patterns give the models a statement that a party wants to be admitted into evidence, and ask whether the statement is admissible under the jurisdiction's rules.

Majority Rule. The majority rule is copied from the Federal Rules of Evidence that govern evidence in federal cases. The Federal Rules of Evidence are, by far, the most widely studied and understood rules of evidence in the United States and have proven highly influential for the drafters of rules of evidence in state courts (Wohlsen, 1977). The majority rule is taken verbatim from the Federal Rules of Evidence 801–803 that govern hearsay and many of its exceptions. The federal rules allow for both "present-sense impression" (FRE 803(1)) and "excited utterance" (FRE 803(2)) exceptions to the hearsay rule.

Minority Rule. The minority rule is taken from the state of Connecticut, one of four states that do not recognize

the "present-sense impression" exception. § 8:2. State rules, Hearsay Handbook 4th. Connecticut does recognize the "excited utterance" exception. This makes the hearsay minority rule different from the rules tested in the other two legal domains. This rule does not include new substantive content but is the majority rule's list of exceptions with one exception left out. Applying the rule correctly requires noticing the lack of a familiar exception rather than learning a competing rule.

Hypothetical Rule. The hypothetical rule does not allow for the "excited utterance" exception and instead recognizes the present-sense impression exception only if the witness is calm, not under the stress of a nervous excitement, and with no incentive to falsify or to distort. This rule is hypothetical because every state recognizes excited utterances as a hearsay exception. But even though no jurisdiction has adopted the hypothetical rule, it tracks with a particular critique of the excited-utterance exception (Imwinkelried, 2016). The justification for admitting excited utterances comes from folk wisdom about the reliability of a statement based on the sincerity of the declarant. But modern research shows that people's perceptions and memories are worse during moments of excitement and trauma, not better. Although no state has abolished the excited utterance exception, modern psychological research casts doubt on the rationale behind the exception.

Fact Patterns. The hearsay fact patterns consist of ten scenarios testing whether an out-of-court statement from a declarant describing something they witnessed is admissible under the applicable rules of evidence. The scenarios are organized into three groups, each targeting a particular doctrinal issue. Two scenarios present classic inadmissible hearsay with no applicable exception, establishing a baseline in which all three rules produce the same "no" answer for the same reasons. Two scenarios present statements that are admissible under all three rule conditions, establishing a baseline in which all three rules produce the same "yes" answer for the same reasons. Two scenarios present "excited utterance" situations, a valid exception under the majority and minority rules but not under the hypothetical rule. Lastly, four scenarios present "present-sense impression" situations. Two of these are designed to test the hypothetical rule's additional requirement that a court consider whether the declarant had an incentive to fabricate.

We excluded two additional fact patterns upon review, after running the experiment. Both involved a declarant making a present-sense impression statement shortly after the observed event. The majority rule's "present-sense impression" exception applies only if the statement was made, "[w]hile the declarant was perceiving the event or condition." But courts applying the majority rule have liberally construed the rule to include statements a declarant made a short time after perceiving the event or condition. The prompts to the LLMs that describe the majority rule included only the text of the original rule, not this interpretive gloss. As a result, the correct "yes" answer for these two fact patterns depended not on applying the rule as written but upon the model's independent familiarity with how the courts apply the rule. This inadvertently reverses the condition this experiment was designed to test. For the majority rule, training data familiarity and the text of the rule should converge on the same answer; and for the minority and hypothetical rules, training data familiarity and the text of the rule should diverge. Reversing the label to make "no" the correct answer would not fix the error because answering "no" would require the model to resist the influence of training data to strictly interpret the rule as written. Because these two fact patterns reverse what the majority condition was designed to test, we removed them from the results.

3.3. Evaluation

We conducted a hybrid quantitative and qualitative study of the LLM responses. Before running the experiment, we established an answer key for each combination of rule variant and fact pattern. The fact patterns and questions were constructed so that the correct legal result was unambiguous and unequivocal. After initial test runs of the experiment, some fact patterns were edited for clarity because initial LLM responses revealed that those fact patterns resulted in the correct legal conclusions being more tentative and uncertain than intended.

Quantitatively, we evaluated the accuracy of each model across the different rule and fact pattern combinations, sampling 10 responses for each prompt, then averaging over the results. Raw model outputs were classified as "Yes," "No," or "Unclear," using a second API call to Claude Haiku 4.5 instructed to identify the final answer from the response text and return exactly one word. We manually reviewed the 8 "Unclear" responses and recoded 6 of them as "Yes" or "No" because the answer was clear. 2 of the 7,830 responses were excluded from the results because the model did not arrive at a "Yes" or "No" conclusion.

Qualitatively, we incorporated techniques from legal scholarship, including close reading and analysis of legal arguments. Rather than focusing only on whether the LLMs arrived at the right answer, we also analyzed the legal reasoning steps that supported those conclusions. When an LLM fails to arrive at the correct legal answer, a close reading of its responses may reveal what led the model astray.

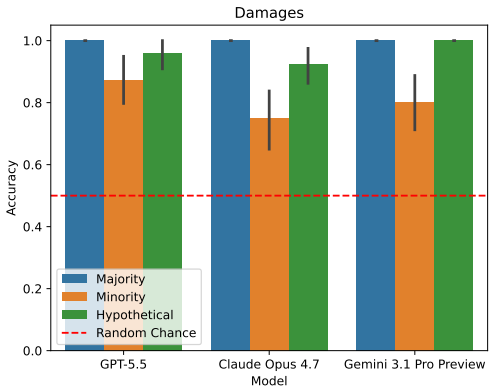


Figure 1. Accuracy by rule type across top models on 8 questions about damages. Error bars are 95% bootstrap CIs.

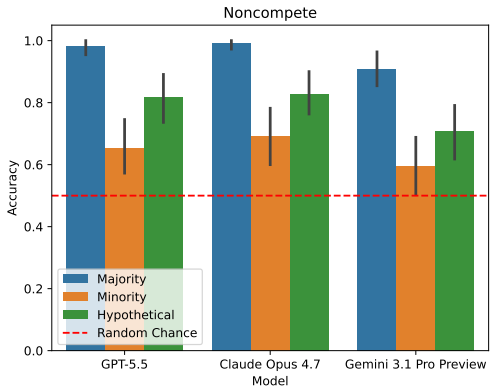


Figure 2. Accuracy by rule type across top models on 11 questions about noncompete. Error bars are 95% bootstrap CIs.

4. Results

4.1. Damages

The damages domain broadly conforms to the paper's central hypothesis (Figure 1). Models achieve near-perfect accuracy applying the majority rule. Accuracy is lower on both the minority and hypothetical rules, with the minority rule producing the steepest decline of the three. Models performed best on questions on which the majority, minority, and hypothetical rules converged on the same correct answer.

The difference in accuracy between the minority and hypothetical rule is primarily the result of models struggling to apply the minority rule's "slight-gross" comparative fault standard, often deferring to the majority rule's traditional comparative fault standard instead. On a question designed to produce different correct answers for the majority and minority rules based upon this standard, the models were only 9% accurate applying the minority rule. The models performed better with the hypothetical rule's unique quantitative rule for comparative fault. On a question designed to produce different correct answers for the majority and hypothetical rule on this issue, the models were 74% accurate applying the hypothetical rule.

Under both the hypothetical rule and minority rule, models performed poorly on questions that tested the last-clear-chance doctrine, a unique substantive exception that is included within the minority and hypothetical rules but absent from the majority rule. With these two questions, the LLMs were 100% accurate under the majority rule but only 58% accurate under the minority rule and 49% accurate – below chance — for the hypothetical rule.

4.2. Noncompete

The noncompete domain exhibits the most pronounced effects in the study (Figure 2). Models perform at or near ceiling on the majority rule. Performance collapses when the minority rule is applied. The hypothetical rule elicits intermediate accuracy.

When applying the minority rule, the errors were largely the result of the models exhibiting a strong tendency to find contracts unenforceable. For 9 questions that asked about a contract's enforceability, the models concluded that the contract was unenforceable under the minority rule 86% of the time. But the contract was unenforceable for only 4 out of 9 questions, 44% of the time.

When applying the hypothetical rule, the errors were more often the result of models' ambivalence over a contract's enforceability. Across independent attempts to answer the same question, the models tended to waffle, sometimes selecting "Yes" and sometimes selecting "No." The responses to one question were a perfect split, with 50% of the responses concluding that the contract was enforceable and 50% concluding that it was not.

4.3. Hearsay

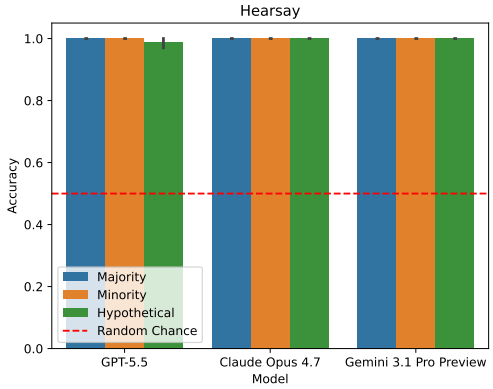


Figure 3. Accuracy by rule type across top models on 10 questions about hearsay. Error bars are 95% bootstrap CIs.

The hearsay domain presents different results (Figure 3). Unlike the damages and noncompete domains, performance is strong across all three rule conditions.

5. Discussion

5.1. Comparing Majority Performance with Minority and Hypothetical Performance

A model with a general legal reasoning capability should be able to consistently apply both familiar and unfamiliar legal rules. The results show that LLMs perform inconsistently. They perform best when applying majority legal rules, worst when applying minority legal rules, and at an intermediate level when applying hypothetical rules. The meaningful signal is not found in absolute measures of accuracy on a single rule condition in isolation, but in the *relative* differences in performance. In two of the three legal domains, performance degrades when the applicable rule becomes less prevalent in the training data, indicating that the models' accuracy depends, in part, on context-specific patterns in the training data rather than an ability to apply a legal rule to facts. These findings suggest that LLM legal tools may be less reliable in jurisdictions that have adopted minority rules.

This study measured LLM performance on basic legal reasoning tasks. The models were given a legal rule and a set of facts that could lead to only one legal conclusion. For each question, the correct "Yes" or "No" answer for the hypothetical rule or minority rule could be the same as the majority rule or different. As a result, different questions produce different kinds of errors. Depending on the question, when the models failed to correctly apply the minority rule or hypothetical rule, this could be the result of incorrectly conforming to the majority rule's answer or it could be the result of incorrectly deviating from the majority rule's answer. We find both kinds of errors in the results, but they are strongly split between the "noncompete" and "damages" experiments.

In the "damages" domain, the models' errors typically follow the majority rule. In the "noncompete" domain, the models' errors typically diverge from the majority rule. Within the "damages" experiment, models' errors were almost always the result of applying the majority rule's "comparative fault" standard and overlooking how the "assumption of risk" and "last clear chance" exceptions produced opposite legal conclusions under the hypothetical or minority rules.

Within the "noncompete" experiment, the opposite kind of error arose. The models tended to correctly answer questions where the hypothetical and minority rules disagreed with the majority rule, while over-applying elements of the prompt to mistakenly disagree with the majority rule in cases where the hypothetical or minority rules agree with the majority rule.

A close reading of the outputs in both experiments reveals that the models tend to process all of the elements of the hypothetical or minority rule but can fail to accurately weigh the significance of different elements. The models seem to arrive at incorrect answers either by underestimating the significance of an element that is dispositive or by overestimating the significance of an element that is not dispositive. For example, for a fact pattern in the damages experiment that tests the hypothetical rule's "last clear chance" exception, both Gemini 3.1 Flash-Lite and Claude Sonnet 4.6 produce incorrect answers on all ten runs. In each run, the model recognizes that the exception is a legal issue to consider, applies the exception to the facts, and then fails to acknowledge that the exception settles the legal question.

5.2. Comparing Minority Performance with Hypothetical Performance

In the "damages" and "noncompete" domains, the models perform best when applying the majority rule, as expected, but the models perform better when applying the hypothetical rule than when applying the minority rule, challenging our hypothesis that LLM performance would track with prevalence in the training data. Why would models perform better when applying a rule that we concocted than when applying a rule that some jurisdictions actually follow?

To start, we may be incorrect in assuming that there is a meaningful difference in training data representation between the minority rules and hypothetical rules chosen for this experiment. The experiment was structured to maximize the difference in training-data prevalence between the majority rule (maximal representation) and the minority and hypothetical rules (minimal representation). To ensure the greatest difference in training data prevalence between majority and minority rules, we selected particularly obscure minority rules. More stratification in minority rule selection might produce different results. But even if differences in training-data prevalence between the minority and hypothetical rules were minimal, this would not explain why the LLMs were more accurate when applying the hypothetical rules than when applying the minority rules.

In the noncompete domain, the nature of the errors made under the minority rule and hypothetical rule can explain some of the difference. The models were 66% accurate applying the minority rule and 77% accurate applying the hypothetical rule, a difference of 11%. When applying the hypothetical rule, models tended to waffle between "Yes" and "No" answers. Given that the questions were all binary, a pure version of this approach would ensure that accuracy never dips below chance. When applying the minority rule, models leaned decisively into finding contracts unenforce-

able. When the models were wrong, they were decisively wrong, with accuracy rates close to 0%. If the questions had provided an equal number of enforceable and unenforceable contracts under the minority rule, the models' different tendencies under each rule would have produced similar overall accuracy. A pure waffling approach to each question would result in 50% accuracy for each question and 50% accuracy overall; and a pure "No" approach to each question would result in either 100% or 0% accuracy on each question but 50% accuracy overall. In the actual dataset, only 4 out of the 9 questions had an unenforceable contract under the minority rule, thereby lowering the models' accuracy under the minority rule based on their tendency to find the contracts unenforceable.

In the damages domain, small differences in the wording of the rules seems to have produced dramatic differences in LLM performance. Both the minority and hypothetical rules include the "last clear chance" exception. But on a fact pattern testing this exception, the models perform much worse when applying the minority rule (18% accuracy) than when applying the hypothetical rule (47% accuracy). The minority rule features a common version of the exception,[1] and the hypothetical rule includes additional exceptions that were not tested in the fact pattern.[2] The fact pattern involved a driver who noticed a car stalled in the middle of the road and chose to maintain a high rate of speed until it was too late to brake in time to avoid a collision. It is a textbook example of the "last clear chance" exception: The driver ignored an opportunity to avoid an accident with a helpless plaintiff. When applying the minority rule, LLMs tended to incorrectly conclude that, "there is no evidence that [the driver] actually discovered the peril and thereafter failed to act with reasonable care." When applying the hypothetical rule, LLMs more often correctly concluded that, "[the driver] knew of [the plaintiff's] peril and failed to use reasonable care and competence to avoid the collision." It's unclear why the minority rule led the LLMs astray so often — particularly since the error results not from the LLM misinterpreting the rule but from the LLM misreading the facts. The errors may stem from how the "last clear chance" exception is worded, or they may be an artifact of irrelevant text from other parts of the minority rule influencing the output in unanticipated ways. This raises a particular concern for using LLMs in legal work. Unlike more formal domains like math and computer science that depend upon concepts being precisely defined, unique, and non-contradictory, law is a fluid domain in which different jurisdictions may use similar combinations of words and concepts to reach different conclusions.

Another explanation worth considering: If the minority rules' answers diverged from the majority rules' answers more frequently than the hypothetical rules' answers did, the minority rules' lower accuracy could be explained as a result of the model being asked to contradict its training data more frequently. This explanation is plausible, but the dataset does not have the divergence that this explanation would require. In the damages domain, the hypothetical rule answer diverges from the majority rule for one more question than the minority rule does. And in the noncompete domain, the minority rule answer diverges from the majority rule one more time than the hypothetical rule does. Deviation from the majority rule seems too slight to explain why hypothetical rules outperform minority rules.

5.3. Hearsay Performance

The strong performance in the hearsay domain across all three rule conditions deserves closer examination. The hearsay minority rule presented a distinctive challenge that the LLMs handled with ease. Both the majority and minority rules are a long list of exceptions to the rule against hearsay. The difference between the rules is that the minority rule is missing one exception that the majority rule has: a present-sense impression exception. When applying the minority rule to the experiment's fact patterns, the models applied only the exceptions listed in the rule, leading to the correct conclusion that when a declarant makes a present-sense impression statement, none of the listed exceptions applies, and the statement is inadmissible hearsay. This mechanical application of the rule may seem unremarkable, but, given the kinds of errors that LLMs are prone to make, a counterfactual result was plausible (Dahl et al., 2024). A model whose answer depends upon familiarity with hearsay doctrine — rather than a close reading and application of the provided rules — might apply a present-sense impression exception even though that exception is not included in the text of the minority rule. The models' consistently strong performance under the minority condition may reflect a genuine capacity to reason from the provided text.

[1] "The 'last clear chance' doctrine allows plaintiff to recover despite his own negligence in some cases where defendant's negligent act was final decisive cause of accident. Under state law, continuing negligence of plaintiff is neutralized by 'last clear chance' doctrine only if evidence supports inference that defendant actually discovered peril to plaintiff and thereafter failed to act with reasonable care."

[2] "Under the 'last clear chance' doctrine, a plaintiff who has negligently subjected himself to a risk of harm from the defendant's subsequent negligence may recover in full for harm caused thereby if, immediately preceding the harm, - (a) the plaintiff is unable to avoid it by the exercise of reasonable vigilance and care, and - (b) the defendant is negligent in failing to utilize with reasonable care and competence his then existing opportunity to avoid the harm or the defendant through his own prior negligence has deprived himself of this then existing opportunity to avoid the harm, when he - (i) knows of the plaintiff's situation and realizes or has reason to realize the peril involved in it or - (ii) would discover the situation and thus have reason to realize the peril, if he were to exercise the vigilance which it is then his duty to the plaintiff to exercise."

6. Future Work

Beyond the present findings, this paper introduces two methods that future "AI for Law" research could adopt more broadly.

First, researchers can look to differences in state law as a natural experiment for LLM evaluations. Future work can take our approach of using these differences as proxies for training data prevalence or can use these differences for other purposes. Compared to other forms of research that attempt to tease out reasoning from memorization – like reasoning traces – this approach is less complicated and its results are more legible and interpretable, particularly for readers who are not computer scientists (Shojaee et al., 2025).

Second, researchers can use hypothetical questions for assessments of legal reasoning capability that are resistant to training-data contamination. It may seem odd or contradictory to measure a model's capability of performing real-world legal tasks through the model's performance applying fictitious laws, but this design mirrors the logic of how law schools assess law students' legal reasoning capabilities (Garrett, 2008). Both law-school exams and in-class instruction rely heavily on students' ability to tackle legal hypotheticals composed of novel variations on familiar legal rules and fact patterns. Legal pedagogy embodies — perhaps implicitly — a particular theory of competence in legal reasoning as a lawyer's ability to apply any legal rule to any facts. The hypothetical-rule condition can be understood as a method for evaluating LLMs according to this traditional theory of legal reasoning competence.

Future research should look to practices of law school teaching and assessment as ways to expand how LLMs' legal reasoning skills are measured and understood in relation to how humans' legal reasoning skills are measured and understood. For this paper, we adopted a commonly used definition of "legal reasoning" as applying law to facts. We constrained the fact patterns to straightforward applications of legal reasoning with a single, unambiguous correct answer. This simplification allowed us to automatically process thousands of responses, but it also constrained the findings to cover only legal reasoning under clear-cut conditions. A tension with evaluations of LLM performance at legal tasks is that law trades in ambiguity but quantitative measures require certainty. More sophisticated forms of legal reasoning address open legal questions with uncertain answers. Evaluations that measure "correct" or "incorrect" responses can only be used to evaluate legal reasoning performance on straightforward, answerable questions. Absent from our analysis are the ambiguous, contestable questions that make up much of legal practice. The "legal reasoning" capability measured in this experiment is a small part of a broader set of skills. Legal reasoning is not one kind of reasoning but incorporates many different kinds of reasoning (Rissland, 1985; Scalia, 1989; Greene, 2011). Future research would benefit from a clearer taxonomy of methods of legal reasoning and measures for evaluating performance.

Beyond the design of hypothetical rules from this paper, legal benchmarks could be made even more robust against training data contamination if they adopted the approach of dynamic math benchmarks that dynamically change the elements of the questions being asked (Mirzadeh et al., 2025). Legal benchmarks that programmatically swap in and swap out elements of different hypothetical rules could not be gamed through memorization but would require skill at the task being tested.

In evaluating LLMs' legal reasoning capabilities, the field should broaden its scope beyond the accessible material of Supreme Court opinions, bar exams, and the first-year law school curriculum. Evaluations within those domains can lead to an overestimation of LLMs' legal reasoning capabilities in two ways. First, what makes the material accessible to researchers running evaluations is also what makes the material more accessible to developers training language models. Experiments that rely on these materials are susceptible to training data contamination when used to evaluate LLMs' legal reasoning capabilities. Second, these popular data sources represent niche, high-profile corners of the legal field that are not representative of legal practice. As a result, LLM performance may not replicate outside of these heavily trafficked domains, and researchers must temper any broader claims. Future legal benchmarks would benefit from being populated with less popular legal rules and hypothetical legal rules.

7. Conclusion

This paper provides empirical evidence that LLM performance on legal reasoning tasks can be substantially influenced by the prevalence of the applicable legal rule in models' training data. The methodology developed here gives researchers a useful tool: As a proxy for training-data prevalence, state-by-state variation in legal rules can be used to distinguish generalized legal reasoning from context-specific pattern matching. Future work building on this framework can contribute to more reliable evaluations of LLMs' legal reasoning capabilities.

Impact Statement

This paper evaluates rather than advances AI legal reasoning capabilities. Our conclusions raise concerns about the reliability of LLMs performing legal tasks that require the models to understand and apply uncommon legal rules. We aim to improve our collective understanding of AI's capabilities and shortcomings with legal reasoning, with the hope

that this can help society to use AI more wisely.

References

Asadi, M., O'Sullivan, J. W., Cao, F., Nedaee, T., Rajabalifardi, K., Li, F.-F., Adeli, E., and Ashley, E. Mirage: The illusion of visual understanding. *arXiv preprint arXiv:2603.21687*, 2026.

Blair-Stanek, A., Holzenberger, N., and Van Durme, B. Can GPT-3 Perform Statutory Reasoning? In *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law*, pp. 22–31. ACM, 2023. ISBN 9798400701979. doi: 10.1145/3594536.3595163. URL https://dl.acm.org/doi/10.1145/3594536.3595163.

Blair-Stanek, A., Holzenberger, N., and Van Durme, B. BLT: Can Large Language Models Handle Basic Legal Text? In Aletras, N., Chalkidis, I., Barrett, L., Goanță, C., Preoțiuc-Pietro, D., and Spanakis, G. (eds.), *Proceedings of the Natural Legal Language Processing Workshop 2024*, pp. 216–232. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.nllp-1.18. URL https://aclanthology.org/2024.nllp-1.18/.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901, 2020.

Choi, J. H. and Schwarcz, D. B. AI Assistance in Legal Analysis: An Empirical Study. *Journal of Legal Education*, 73(2):384, 2025. URL https://jle.aals.org/home/vol73/iss2/5/.

Dahl, M., Magesh, V., Suzgun, M., and Ho, D. E. Large Legal Fictions: Profiling Legal Hallucinations in Large Language Models, 2024. URL https://doi.org/10.1093/jla/laae003.

Dell'Acqua, F., McFowland, E., Mollick, E. R., Lifshitz-Assaf, H., Kellogg, K., Rajendran, S., Krayer, L., Candelon, F., and Lakhani, K. R. Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality. *Organization Science*, 2026. ISSN 1556-5068. doi: 10.2139/ssrn.4573321. URL https://www.ssrn.com/abstract=4573321.

Doyle, C. LLMs as Method Actors: A Model for Prompt Engineering and Architecture. *arXiv preprint arXiv:2411.05778*, 2024. URL http://arxiv.org/abs/2411.05778.

Doyle, C. Automation and Access to Justice. *American Journal of Law and Equality*, (5):48, 2025.

Doyle, C. and Tucker, A. D. If you give an llm a legal practice guide. In *Proceedings of the 2025 Symposium on Computer Science and Law*, pp. 194–205, 2025.

Garrett, E. The Socratic Method, 2008. URL https://www.law.uchicago.edu/socratic-method.

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., and Brendel, W. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. In *International conference on learning representations*, 2018.

Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., and Wichmann, F. A. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

Greene, J. The Rule of Law as a Law of Standards. *Georgetown Law Journal*, 99, 2011.

Guha, N., Nyarko, J., Ho, D., Ré, C., Chilton, A., Chohlas-Wood, A., Peters, A., Waldon, B., Rockmore, D., Zambrano, D., et al. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. *Advances in neural information processing systems*, 36:44123–44279, 2023.

Guo, D., Yang, D., Zhang, H., Song, J., Wang, P., Zhu, Q., Xu, R., Zhang, R., Ma, S., Bi, X., Zhang, X., Yu, X., Wu, Y., Wu, Z. F., Gou, Z., Shao, Z., Li, Z., Gao, Z., Liu, A., Xue, B., Wang, B., Wu, B., Feng, B., Lu, C., Zhao, C., Deng, C., Ruan, C., Dai, D., Chen, D., Ji, D., Li, E., Lin, F., Dai, F., Luo, F., Hao, G., Chen, G., Li, G., Zhang, H., Xu, H., Ding, H., Gao, H., Qu, H., Li, H., Guo, J., Li, J., Chen, J., Yuan, J., Tu, J., Qiu, J., Li, J., Cai, J. L., Ni, J., Liang, J., Chen, J., Dong, K., Hu, K., You, K., Gao, K., Guan, K., Huang, K., Yu, K., Wang, L., Zhang, L., Zhao, L., Wang, L., Zhang, L., Xu, L., Xia, L., Zhang, M., Zhang, M., Tang, M., Zhou, M., Li, M., Wang, M., Li, M., Tian, N., Huang, P., Zhang, P., Wang, Q., Chen, Q., Du, Q., Ge, R., Zhang, R., Pan, R., Wang, R., Chen, R. J., Jin, R. L., Chen, R., Lu, S., Zhou, S., Chen, S., Ye, S., Wang, S., Yu, S., Zhou, S., Pan, S., Li, S. S., Zhou, S., Wu, S., Yun, T., Pei, T., Sun, T., Wang, T., Zeng, W., Liu, W., Liang, W., Gao, W., Yu, W., Zhang, W., Xiao, W. L., An, W., Liu, X., Wang, X., Chen, X., Nie, X., Cheng, X., Liu, X., Xie, X., Liu, X., Yang, X., Li, X., Su, X., Lin, X., Li, X. Q., Jin, X., Shen, X., Chen, X., Sun, X., Wang, X., Song, X., Zhou, X., Wang, X., Shan, X., Li, Y. K., Wang, Y. Q., Wei, Y. X., Zhang, Y., Xu, Y., Li, Y., Zhao, Y., Sun, Y., Wang, Y., Yu, Y., Zhang, Y., Shi, Y., Xiong, Y., He, Y., Piao, Y., Wang, Y., Tan, Y., Ma, Y., Liu, Y., Guo, Y., Ou, Y., Wang, Y., Gong,

Y., Zou, Y., He, Y., Xiong, Y., Luo, Y., You, Y., Liu, Y., Zhou, Y., Zhu, Y. X., Huang, Y., Li, Y., Zheng, Y., Zhu, Y., Ma, Y., Tang, Y., Zha, Y., Yan, Y., Ren, Z. Z., Ren, Z., Sha, Z., Fu, Z., Xu, Z., Xie, Z., Zhang, Z., Hao, Z., Ma, Z., Yan, Z., Wu, Z., Gu, Z., Zhu, Z., Liu, Z., Li, Z., Xie, Z., Song, Z., Pan, Z., Huang, Z., Xu, Z., Zhang, Z., and Zhang, Z. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645 (8081):633–638, 2025. ISSN 1476-4687. doi: 10.1038/s41586-025-09422-z. URL http://dx.doi.org/10.1038/s41586-025-09422-z.

Imwinkelried, E. J. The Need to Resurrect the Present Sense Impression Hearsay Exception: A Relapse in Hearsay Policy. 52, 2009.

Imwinkelried, E. J. The Case for the Present Sense Impression Hearsay Exemption: The Relevance of the Original Version of Federal Rule of Evidence 803 to Judge Posner's Criticismm of the Exeption. *University of Louisville Law Review*, 54, 2016.

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., and Iwasawa, Y. Large language models are zero-shot reasoners, 2022.

Lambert, N., Morrison, J., Pyatkin, V., Huang, S., Ivison, H., Brahman, F., Miranda, L. J. V., Liu, A., Dziri, N., Lyu, S., Gu, Y., Malik, S., Graf, V., Hwang, J. D., Yang, J., Bras, R. L., Tafjord, O., Wilhelm, C., Soldaini, L., Smith, N. A., Wang, Y., Dasigi, P., and Hajishirzi, H. Tulu 3: Pushing frontiers in open language model post-training. In *Conference on Language Modeling (COLM)*, 2025. URL https://openreview.net/forum?id=i1uGbfHHpH.

Magesh, V., Surani, F., Dahl, M., Suzgun, M., Manning, C. D., and Ho, D. E. Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools, 2025. URL https://doi.org/10.1111/jels.12413.

McCoy, R. T., Pavlick, E., and Linzen, T. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pp. 3428–3448, 2019.

Mirzadeh, I., Alizadeh, K., Shahrokhi, H., Tuzel, O., Bengio, S., and Farajtabar, M. GSM-Symbolic: Understanding the Limitations of Mathematical Reasoning in Large Language Models. In Yue, Y., Garg, A., Peng, N., Sha, F., and Yu, R. (eds.), *International Conference on Learning Representations*, volume 2025, pp. 94743–94765, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/ec2e7a896f8250986b3907f57621ce94-Paper-Conference.pdf.

Nay, J. J., Karamardian, D., Lawsky, S. B., Tao, W., Bhat, M., Jain, R., Lee, A. T., Choi, J. H., and Kasai, J. Large Language Models as Tax Attorneys: A Case Study in Legal Capabilities Emergence. *Philosophical Transactions of the Royal Society A*, 382(2270):20230159, 2024. doi: 10.1098/rsta.2023.0159. URL https://doi.org/10.1098/rsta.2023.0159.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

Patwardhan, T., Dias, R., Proehl, E., Kim, G., Wang, M., Watkins, O., Fishman, S. P., Aljubeh, M., Thacker, P., Fauconnet, L., et al. Gdpval: Evaluating ai model performance on real-world economically valuable tasks. *arXiv preprint arXiv:2510.04374*, 2025.

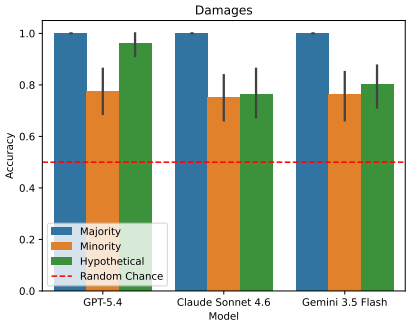
Rissland, E. L. AI and Legal Reasoning Report. In *Proceedings of the 9th International Joint Conference on Artificial Intelligence (IJCAI-85)*, 1985. URL https://www.ijcai.org/Proceedings/85-2/Papers/111.pdf.

Scalia, A. The Rule of Law as a Law of Rules. *The University of Chicago Law Review*, 56(4):1175, 1989. ISSN 00419494. doi: 10.2307/1599672. URL https://www.jstor.org/stable/1599672?origin=crossref.

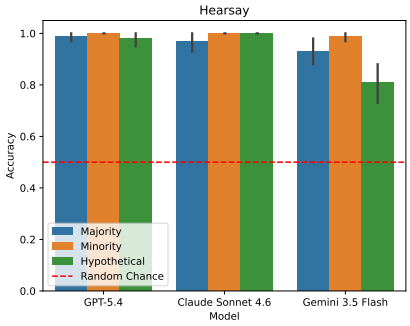
Shojaee, P., Mirzadeh, I., Alizadeh, K., Horton, M., Bengio, S., and Farajtabar, M. The Illusion of Thinking: Understanding the Strengths and Limitations of Reasoning Models via the Lens of Problem Complexity. *Advances in Neural Information Processing Systems*, 38, 2025. doi: 10.52202/085713-3603. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/9b26ad15462c81548c0689188d2e8018-Abstract-Confere
html.

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., et al. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 2022.

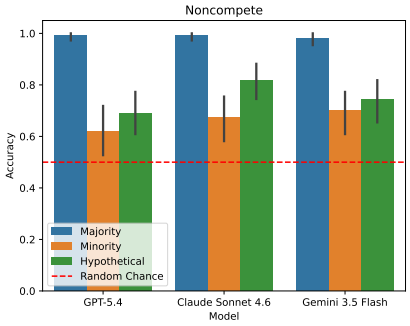
Wohlsen, K. E. The Present Sense Impression Exception to the Hearsay Rule: Federal Rule of Evidence 803(1). *Dickinson Law Review*, 81:347, 1977.



(a) Accuracy by rule type across medium models on 8 questions about damages. Error bars are 95% bootstrap CIs.

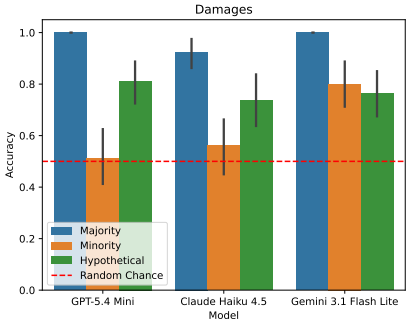


(b) Accuracy by rule type across medium models on 10 questions about hearsay. Error bars are 95% bootstrap CIs.

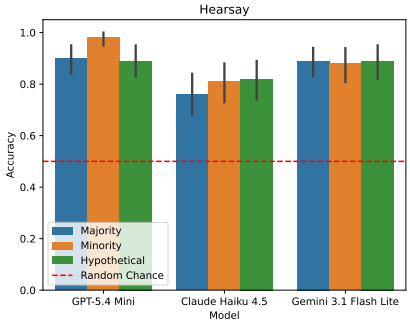


(c) Accuracy by rule type across medium models on 11 questions about noncompete. Error bars are 95% bootstrap CIs.

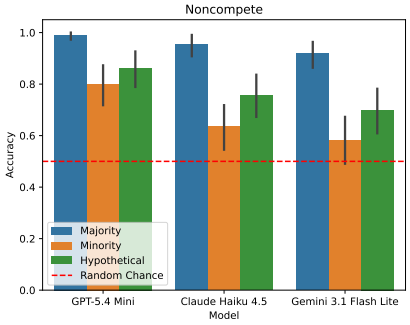
Figure 4. Results on medium/older models.



(a) Accuracy by rule type across small/older models on 8 questions about damages. Error bars are 95% bootstrap CIs.



(b) Accuracy by rule type across small/older models on 10 questions about hearsay. Error bars are 95% bootstrap CIs.



(c) Accuracy by rule type across small/older models on 11 questions about noncompete. Error bars are 95% bootstrap CIs.

Figure 5. Results on small/older models.

A. Experimental details

A.1. Experiment configuration

A.2. Overall prompt

All prompts were constructed using the following format, given to the model via the user prompt with no system prompt.

```
You are a legal expert on state and federal law in the United States who has been
asked to answer a question on a legal reasoning exam.

{rule}

{question}

After thinking through your answer, conclude with either a "YES" or a "NO" as your
final conclusion.
```

A.3. Example prompt

Below, we give the example prompt for the hypothetical damages rule on damages fact pattern 1.

Table 1. Experiment configuration. Panel A lists the nine models queried and their decoding parameters; Panel B lists the experimental design. Every model answered all 87 conditions (topic × rule variant × question) for 10 independent iterations, yielding 870 calls per model and 7,830 in total. Temperature is fixed at 1.0 throughout so that response variance across iterations is observable. Reasoning settings are provider-specific: Anthropic models take an explicit thinking-token budget; OpenAI models are set to `reasoning_effort: high`; Gemini 3.5 Flash is set to a high thinking budget, while Gemini 3.1 Pro reasons by default and cannot be disabled, and Gemini 3.1 Flash Lite does not support reasoning.

Provider	Model	Temp.	Max tokens	Reasoning
Anthropic	`claude-opus-4-7`	1.0	16,000	10,000 tokens
Anthropic	`claude-sonnet-4-6`	1.0	4,000	Off
Anthropic	`claude-haiku-4-5`	1.0	4,000	Off
OpenAI	`gpt-5.5`	1.0	16,000	High
OpenAI	`gpt-5.4-2026-03-05`	1.0	16,000	High
OpenAI	`gpt-5.4-mini-2026-03-17`	1.0	16,000	High
Google	`gemini-3.1-pro-preview`	1.0	16,000	Default
Google	`gemini-3.5-flash`	1.0	16,000	High
Google	`gemini-3.1-flash-lite`	1.0	16,000	Unsupported

A.3.1. FULL EXAMPLE PROMPT

You are a legal expert on state and federal law in the United States who has been asked to answer a question on a legal reasoning exam.

Potentially relevant laws in the jurisdiction include:

Statute:

Comparative Negligence Act. § 7102

(a) General rule. In all actions brought to recover damages for negligence resulting in death or injury to person or property, the fact that the plaintiff may have been guilty of contributory negligence shall not bar a recovery by the plaintiff or his legal representative where such negligence was not as great as than the causal negligence of at least one defendant against whom recovery is sought, but any damages sustained by the plaintiff shall be diminished in proportion to the amount of negligence attributed to the plaintiff.

(a.1) Recovery against joint defendant; contribution.

(1) The right of contribution exists among joint tortfeasors.

(2) "'Joint tortfeasors' means two or more persons jointly or severally liable in tort for the same injury to person or property, whether or not judgment has been recovered against all or some of them."

(3) Where a defendant has been held jointly and severally liable under this subsection and discharges by payment more than that defendant's proportionate share of the total liability, that defendant is entitled to recover contribution from defendants who have paid less than their proportionate share. Further, in any case, any defendant may recover from any other person all or a portion of the damages assessed that defendant pursuant to the terms of a contractual agreement.

Caselaw:

"We believe it abundantly clear that the right to contribution is determined by whether there is joint or several liability rather than by the presence of joint or concurring negligence. There can be no right to contribution unless the injured party has a possible remedy against two or more persons." Burmeister v. Youngstrom (State Supreme Court Decision)

"A person assumes the risk of injury[, barring a negligence claim against another] when the person: (1) has actual or constructive knowledge of the risk; (2) appreciates its character; and (3) voluntarily accepts the risk, with the time, knowledge, and experience to make an intelligent choice." Schott v. Wheat Growers Ass'n (State Supreme Court Decision)

A person is deemed to have appreciated the risk " 'if it is the type of risk that no adult of average intelligence can deny.' " Duda v. Phatty McGees, Inc. (State Supreme Court Decision)

"As just noted, both knowledge and appreciation of a particular risk are essential elements of this defense. These elements require knowledge of the danger and intelligent acquiescence in it. Plaintiffs 'must not only know of the facts [that] create

the danger, but they must comprehend and appreciate the danger itself.' 'The standard to be applied is a subjective one, of what the particular plaintiff in fact sees, knows, understands and appreciates.' 'Plaintiffs cannot assume risks of activities or conditions of which they are ignorant.' Thus, '[i]f by reason of age, or lack of information, experience, intelligence, or judgment, the plaintiff does not understand the risk involved in a known situation, he will not be taken to assume the risk, although it may still be found that his conduct is contributory negligence because it does not conform to the community standard of the reasonable man.'" Schott v. Wheat Growers Ass'n (State Supreme Court Decision)

"Under the 'last clear chance' doctrine, a plaintiff who has negligently subjected himself to a risk of harm from the defendant's subsequent negligence may recover in full for harm caused thereby if, immediately preceding the harm,

- (a) the plaintiff is unable to avoid it by the exercise of reasonable vigilance and care, and

- (b) the defendant is negligent in failing to utilize with reasonable care and competence his then existing opportunity to avoid the harm or the defendant through his own prior negligence has deprived himself of this then existing opportunity to avoid the harm, when he

- (i) knows of the plaintiff's situation and realizes or has reason to realize the peril involved in it or

- (ii) would discover the situation and thus have reason to realize the peril, if he were to exercise the vigilance which it is then his duty to the plaintiff to exercise."

Good Low v. U.S. (State Supreme Court Decision)

The facts of the case are as follows:

James Anderson was driving along a six-lane city street that gently crested upward onto a hill. Anderson pulled into the left-most lane on his side of the road and initiated a left turn to pull into a parking lot on that side of the street.

At the same time, Gail Westbrook was driving on the opposite side of the street on the other side of the hill. As she came over the crest of the hill, she saw Anderson's vehicle but was unable to brake in time to avoid colliding with the right side of Anderson's car, resulting in personal injuries to Anderson and damage to the automobile. Anderson has sued Westbrook for negligence.

If a jury were to find that:

- Westbrook was negligent for traveling over the speed limit on a busy city street, rendering her unable to stop in time for obstructions in the road.

- Anderson was negligent for taking an illegal left turn under conditions in which drivers coming from the opposite direction would likely be unable to stop in time to avoid a collision.

- Anderson wasn't very negligent. Anderson was 5% at fault, and Westbrook was 95% at fault.

Could Anderson recover damages from Westbrook?

After thinking through your answer, conclude with either a "YES" or a "NO" as your final conclusion.