Demographic Sensitivity in Legal-Rule Retrieval by Large Language Models

Shahzada Muhammad Ali[a,*], Nazia Bibi[a], Faisal Shafait[b] and Adnan Ul Hasan[b]

[a]Department of Computer Software Engineering, College of Signals, National University of Sciences and Technology, Islamabad, Pakistan
[b]Department of Computer Science, School of Electrical Engineering and Computer Science, National University of Sciences and Technology, Islamabad, Pakistan

ARTICLE INFO

Keywords:
large language models
legal reasoning
legal-rule retrieval
demographic sensitivity
algorithmic bias
fairness

ABSTRACT

Large language models (LLMs) are increasingly applied to legal reasoning tasks, yet existing evaluations primarily assess whether models produce correct final answers rather than whether their retrieval of underlying legal rules remains stable when legally irrelevant information is introduced. This paper investigates whether demographic and social descriptors that have no bearing on the applicable law affect the legal rules retrieved by LLMs from otherwise identical legal fact patterns. Using original Multistate Bar Examination (MBE) fact patterns as a baseline, we construct controlled demographic and social variants while preserving the underlying legally relevant facts. Each variant is evaluated across five repeated generations to account for variability in model outputs. We compare the legal rules retrieved by the model against the ground-truth rules associated with the original fact patterns using F1 scores. The results reveal substantial variation in legal-rule retrieval across demographic variants and across repeated generations of the same variant. These differences occur despite the demographic descriptors being introduced without altering the legally relevant facts. The findings indicate that LLM-based legal-rule retrieval may be sensitive to demographic and social information that should be legally irrelevant to the underlying case. This has implications for the evaluation, auditing, and deployment of LLMs in legal contexts, particularly with respect to the stability, consistency, and reliability of automated legal reasoning.

1. Introduction

Large language models (LLMs) are increasingly evaluated on legal tasks requiring the identification of factual elements and issues [18, 26], the retrieval of primary law and precedents [59, 80], and the application of legal rules to complex scenarios [11, 16, 69]. Beyond standard benchmark tasks, real-world deployments of automated decision systems in public governance have repeatedly demonstrated severe societal risks, including discriminatory profiling in immigration and asylum assessments [20, 50], arbitrary denials in welfare eligibility systems [25], and compounded racial bias within criminal justice pipelines [82]. These documented harms have motivated an extensive interdisciplinary literature examining how mathematical fairness metrics align with statutory non-discrimination frameworks [1, 23, 28, 30, 71, 72, 73]. Nevertheless, significant legal, evidentiary, and operational gaps remain in translating technical bias mitigation into practical compliance [1, 20, 28, 71, 82].

Practitioners, researchers, and model developers continue to lack precise guidance on whether the legal reasoning of these systems remains stable when a case contains information that is legally irrelevant to the question presented. This issue is particularly important in legal applications, where seemingly incidental characteristics of the parties should not alter the legal rules that govern an otherwise identical set of facts. This paper investigates whether LLMs exhibit such invariance, and identifies a potentially important source of instability: the introduction of demographic and social descriptors that have no bearing on the applicable law.

First, existing evaluations of LLMs in legal reasoning have largely focused on whether models arrive at the correct legal answer, rather than whether the intermediate legal rules, doctrines, and concepts retrieved by the model remain stable when legally irrelevant information is introduced [4, 22, 26, 32, 46, 47, 54, 60]. Conventional benchmarks typically evaluate a model against a fixed question and fact pattern, measuring whether its final answer corresponds to a known correct answer [19, 21, 34, 53, 55, 76]. While such evaluations are useful for assessing overall performance, they provide limited insight into whether a model has identified the correct legal principles for the right reasons, or whether its legal reasoning is sensitive to characteristics of the individuals described in the case. A model may therefore achieve the correct answer while nevertheless retrieving different legal rules when the underlying legally relevant facts remain unchanged.

Second, existing approaches to evaluating demographic effects in language models frequently treat changes in model outputs as evidence of bias without isolating whether the changed information is legally relevant to the task [29, 48, 61, 74]. In legal reasoning, this distinction is critical. A change in output following the introduction of a demographic characteristic is not necessarily problematic if that characteristic has legitimate legal relevance. Conversely, where the characteristic has no bearing on the governing

*Corresponding author
sali.mscs24mcs@student.nust.edu.pk (S.M. Ali); nazia.bibi@mcs.edu.pk (N. Bibi); faisal.shafait@seecs.edu.pk (F. Shafait); adnan.ulhassan@seecs.edu.pk (A.U. Hasan)
ORCID(s): 0009-0008-8285-390X (S.M. Ali)

law, changes in the legal rules retrieved by the model are difficult to justify on legal grounds. The relevant question is therefore not simply whether models respond differently to different demographic groups, but whether they alter their identification of applicable legal rules in response to information that should be legally inert.

Third, there is a methodological gap in existing evaluations of LLM-based legal reasoning concerning generation variability. LLM outputs are not necessarily deterministic: repeated generations of the same prompt may produce different reasoning, authorities, or conclusions. Consequently, a single observation for each fact pattern may obscure substantial variation in model behaviour. This is particularly important when evaluating demographic sensitivity. If a demographic variant produces a different legal rule in one generation but not another, it is necessary to distinguish isolated generation-level variation from a systematic difference associated with the demographic descriptor. Repeated evaluation of otherwise identical inputs is therefore necessary to assess the stability and reproducibility of observed effects.

Fourth, comparatively little attention has been given to evaluating demographic sensitivity at the level of legal rule retrieval itself. Much of the existing discussion surrounding bias in LLMs concerns differences in generated language, sentiment, toxicity, classification, or downstream decisions [2, 14, 17, 31, 35, 37, 44]. Legal reasoning presents a distinct problem. The output of interest is not merely the model's description of a person or its ultimate prediction, but the set of legal rules, doctrines, and concepts that the model determines to be relevant to a particular fact pattern. If demographic information that is legally irrelevant changes this set of retrieved legal rules, then the effect occurs at an earlier stage of legal reasoning and may subsequently influence the model's final analysis or conclusion.

This paper addresses these gaps by treating legally irrelevant demographic variation as a controlled perturbation of otherwise identical legal fact patterns. Using original MBE fact patterns – Multistate Bar Examination fact patterns, a standard testbed for evaluating legal reasoning – as a baseline, we construct ten demographic and social variants of each fact pattern by introducing descriptors such as "black," "white," "Arab," "female," "homeless," "immigrant," and "wealthy," while preserving the underlying legally relevant facts. Each variant is then presented to the LLM across five repeated generations to account for generation-level variability. The legal rules identified by the model are compared against the ground-truth rules associated with the original fact pattern.

The results reveal substantial variation in legal-rule retrieval across demographic variants despite the underlying legal facts remaining constant. Model performance differed not only across demographic conditions but also across repeated generations of the same condition. These differences are notable precisely because the demographic descriptors were introduced without changing the facts that determine the applicable law.

These findings suggest that legal-rule retrieval by LLMs may not be invariant to demographic and social information that is legally irrelevant to the underlying case. More broadly, they raise questions about whether apparent legal reasoning by LLMs is determined solely by legally salient features of a fact pattern, or whether models may also condition their retrieval of legal knowledge on incidental characteristics of the people involved. This distinction has important implications for the evaluation, deployment, and auditing of LLMs in legal contexts, where consistency with legally relevant facts is a fundamental requirement of reliable legal analysis.

Contributions

To address these gaps, this paper makes four core contributions at the intersection of LLM evaluation, legal reasoning, and demographic sensitivity.

1. We introduce a controlled methodology for testing whether legally irrelevant demographic and social descriptors affect the legal rules, doctrines, and concepts retrieved by an LLM from otherwise identical legal fact patterns.
2. We distinguish demographic sensitivity in legal-rule retrieval from conventional measures of demographic bias, enabling us to determine whether changes in retrieved legal rules arise from legally relevant differences in the facts or from demographic information that should be legally irrelevant.
3. We incorporate repeated generations for each demographic variant, allowing observed differences in legal-rule retrieval to be evaluated alongside generation-level variability rather than treating a single model output as representative of model behavior.
4. We provide empirical evidence that demographic and social descriptors can be associated with substantial differences in the legal rules retrieved by an LLM, raising broader questions about the stability, consistency, and reliability of LLM-based legal reasoning.

The remainder of this paper is structured as follows. Section 2 reviews related work on LLM-based legal reasoning, demographic sensitivity, algorithmic bias, and the evaluation of legal reasoning systems. Section 3 presents our approach for evaluating the effect of legally irrelevant demographic and social descriptors on legal-rule retrieval, including the construction of controlled variants and the use of repeated generations to account for model variability. Section 4 describes the implementation of our evaluation, including the selection of MBE fact patterns, demographic and social perturbations, model prompting, repeated generations, and the metrics used to compare retrieved legal rules against the ground truth. Section 5 presents and discusses the results, examining differences in legal-rule retrieval across demographic variants and repeated generations and considering their implications for the stability and reliability of LLM-based legal reasoning. Section 6 concludes the paper

by summarizing the main findings, discussing their implications and limitations, and outlining directions for future research.

2. Related Work

2.1. LLMs for Legal Reasoning

The emergence of large language models (LLMs) has led to extensive investigation of their ability to perform tasks requiring legal knowledge and reasoning. Early evaluations demonstrated that general-purpose LLMs could achieve substantial performance on established legal examinations and question-answering tasks, challenging the assumption that effective legal reasoning necessarily requires models specifically trained on legal data. Katz et al. [33] provided an influential demonstration by evaluating GPT-4 on the Uniform Bar Examination (UBE), including the Multistate Bar Examination (MBE), Multistate Essay Examination (MEE), and Multistate Performance Test (MPT). GPT-4 achieved 75.7% accuracy on the MBE and an estimated composite UBE score of approximately 297, exceeding the passing threshold across UBE jurisdictions and corresponding to a relatively high percentile among human examinees. These results provided evidence that general-purpose LLMs can acquire and apply substantial amounts of legal knowledge across multiple areas of law, including civil procedure, constitutional law, contracts, criminal law and procedure, evidence, real property, and torts.

Subsequent work has examined the extent to which these results depend on evaluation methodology and prompting conditions. Martínez [40] re-evaluated GPT-4's bar examination performance and successfully replicated the reported MBE accuracy, obtaining 75.6% compared with 75.7% in the original evaluation. However, the study also demonstrated that reported performance can vary substantially depending on the reference population and prompting strategy. In particular, tailored few-shot prompting increased MBE accuracy to 79.5%, whereas minimal prompting produced considerably lower performance. These findings indicate that aggregate benchmark scores do not necessarily represent a fixed level of legal reasoning ability, but can depend on how the legal task is presented and evaluated. More recent large-scale benchmarking by Stubenberg et al. [63] further illustrates the progress of LLMs on standardized legal examinations. Evaluating 52 models from seven major providers on 210 MBE questions, the study found that three models achieved accuracies above 94%, with OpenAI's o3-2025-04-16 achieving the highest score of 96.7% (203 of 210 questions). At the same time, performance varied substantially across the seven legal subject areas, with models generally performing best on Constitutional Law but differing in which categories they handled most successfully. These results suggest that aggregate MBE accuracy can mask meaningful differences in performance across specific areas of legal knowledge.

Beyond standardized examinations, researchers have developed broader benchmarks to assess the legal knowledge and reasoning capabilities of LLMs across a wider range of tasks. LEGALBENCH, introduced by Guha et al. [26], provides a collaboratively constructed benchmark containing a diverse collection of legal reasoning tasks designed to evaluate LLMs across different aspects of legal analysis rather than relying on a single examination format. Similarly, LawBench by Fei et al. [18] evaluates LLMs on a broad set of legal tasks intended to measure legal knowledge and capabilities across different dimensions of legal understanding. These benchmarks represent an important shift from evaluating whether a model can answer a limited set of standardized examination questions toward evaluating legal capabilities across heterogeneous tasks and jurisdictions.

More recent benchmark research has increasingly focused on aspects of legal reasoning that are not captured by conventional multiple-choice accuracy. LEXAM [16], for example, evaluates LLMs on 340 law examinations and extends legal assessment beyond a narrow set of standardized benchmark questions. GREEKBARBENCH [7] similarly evaluates free-text legal reasoning and citation capabilities, highlighting the importance of assessing not only whether a model selects the correct answer but also whether it can produce legally appropriate reasoning and supporting authorities. Research by Zheng et al. [80] introduces a reasoning-focused legal retrieval benchmark, further emphasizing retrieval as an important component of legal reasoning. This line of work is particularly relevant because legal analysis depends not only on producing a plausible conclusion, but also on identifying and applying the legal authorities, rules, and doctrines that govern the facts of a case.

The literature also demonstrates that strong performance on legal questions does not necessarily imply reliable legal reasoning in realistic settings. Magesh et al. [39], for example, evaluated leading AI-powered legal research tools and examined the reliability of their generated legal research outputs, drawing attention to problems such as hallucinated or unsupported legal information. The broader survey literature likewise identifies hallucination, demographic bias, jurisdictional brittleness, and limited explainability as continuing challenges for the deployment of LLMs in legal systems (Dehghani et al. [13]). These concerns are important because legal reasoning requires more than producing an answer that appears plausible: the answer must be grounded in the applicable legal framework and supported by accurate legal knowledge.

Taken together, this literature establishes that LLMs have developed substantial capabilities in legal question answering, examination performance, legal knowledge assessment, and increasingly sophisticated legal reasoning and retrieval tasks. However, the dominant evaluation paradigm remains oriented toward the correctness or quality of the model's final output. Even where benchmarks incorporate explanations, citations, free-text reasoning, or retrieval, the central evaluation question is generally whether the model performs the target task successfully under the evaluated input conditions. The MBE studies are particularly illustrative: Katz et al. [33] and Stubenberg et al. [63] evaluate whether the model selects the correct answer, while

Martínez [40] demonstrates that the measured performance itself can change with prompting and evaluation conditions.

Taken together, this literature establishes that LLMs possess substantial capabilities in legal question answering, examination performance, legal knowledge assessment, and increasingly sophisticated legal reasoning and retrieval. However, existing evaluations remain predominantly focused on the correctness or quality of the model's final output. Even when benchmarks incorporate explanations, citations, free-text reasoning, or retrieval, they generally assess whether the model performs the target task successfully under a given input condition. The literature therefore provides limited evidence about the stability of the legal rules identified during reasoning, particularly when the legally relevant facts remain unchanged. A key unresolved question is whether LLMs identify the same applicable legal rules when legally irrelevant demographic or social descriptors are introduced. The next section therefore examines recent work that evaluates the intermediate processes underlying LLM-based legal reasoning, with particular attention to legal-rule identification and retrieval as components that may reveal limitations not captured by final-answer accuracy.

2.2. Evaluating the Legal Reasoning Process

As LLMs have demonstrated increasingly strong performance on legal question answering and examination benchmarks, recent research has questioned whether final-answer accuracy is sufficient to establish that a model has performed reliable legal reasoning. Legal reasoning is not limited to selecting a correct conclusion; it involves identifying the legally relevant issues, recalling the governing rules, applying those rules to the facts, and deriving a conclusion that is supported by the applicable legal framework. Consequently, several recent studies have begun to evaluate intermediate components of legal reasoning, including rule identification, rule application, reasoning-chain soundness, and the retrieval of relevant legal authorities.

One important development is the decomposition of legal reasoning into distinct stages. Guha et al. [26] introduced LEGALBENCH, which evaluates six categories of legal reasoning: issue-spotting, rule-recall, rule-application, rule-conclusion, interpretation, and rhetorical understanding. Rather than treating legal reasoning as a single question-answering task, LEGALBENCH therefore distinguishes between identifying the relevant legal issue and rule, applying that rule to a factual situation, and reaching a conclusion. The benchmark also evaluates explanations generated for rule-application tasks along separate dimensions of correctness and analysis, where analysis captures whether the model articulates the necessary legal inferences connecting facts to legal conclusions. The results show that even relatively capable models can perform differently across these components, with multi-step rule application and conclusion tasks remaining more challenging than simpler forms of legal information processing. The authors further demonstrate sensitivity to prompt formulation and in-context demonstrations, indicating that measured legal reasoning performance can depend on how the reasoning task is presented.

Dai et al. [11] similarly decompose legal reasoning according to the structure of the legal syllogism. Their LAiW benchmark distinguishes Fundamental Information Retrieval, Legal Principles Inference, and Advanced Legal Application. Of particular relevance, Fundamental Information Retrieval includes legal article recommendation and the identification of factual and legal elements, thereby treating retrieval of the governing legal provision as a distinct capability rather than merely an implicit component of answer generation. Their results indicate that strong surface-level generation does not necessarily correspond to reliable performance on foundational legal reasoning capabilities: several legal LLMs performed comparatively well on advanced generation tasks while struggling with fundamental legal article and element identification. Human evaluation further showed that fluent outputs do not necessarily provide accurate legal reasoning or correct legal citations.

More fine-grained evaluation has examined the internal quality of the reasoning chains produced by LLMs. Mishra et al. [43] developed an error taxonomy for step-by-step legal reasoning that distinguishes premise-level errors, such as misinterpretation, irrelevant premises, and factual hallucinations, from errors occurring when conclusions are derived from incorrect or incomplete premises. They evaluated reasoning chains generated for U.S. Civil Procedure multiple-choice questions using both final-answer accuracy and step-level measures of soundness and correctness. A central finding was that final-answer accuracy substantially overstated reasoning quality: models could produce a correct answer while relying on false premises. In particular, the study identified instances of "Correct Conclusion from False Premises," demonstrating that a correct final answer does not necessarily indicate that the underlying legal reasoning is sound.

This distinction is important because it exposes a limitation of evaluating legal reasoning solely through the endpoint of the reasoning process. If a model reaches the correct conclusion despite misidentifying a legal rule or relying on an incorrect premise, conventional accuracy measures will classify the response as successful. Such a model may therefore appear legally competent while possessing an unstable or erroneous representation of the governing law. Evaluating intermediate reasoning components can reveal these failures, but it also raises a further question: which intermediate component should be examined when the objective is to understand how an LLM selects the law that governs a case?

Recent work has increasingly treated legal knowledge retrieval as one such component. Louis et al. [38] investigate long-form statutory legal question answering using a retrieve-then-read architecture in which a dense retriever first identifies relevant statutory provisions and an LLM subsequently generates an answer based on the retrieved material. Their results demonstrate that retrieval quality remains a significant bottleneck: although the domain-specific

retriever achieved substantially better retrieval performance than general baselines, incomplete retrieval continued to affect downstream answer generation. Their evaluation also found extremely poor performance in extracting the statutory paragraphs that supported generated answers, together with instances of factual and legal hallucination. These findings demonstrate that a model can generate fluent legal text while failing to retrieve or faithfully identify the legal provisions required to support that text.

Paul et al. [52] examine this retrieval problem more directly through Legal Statute Identification, treating identification of the relevant statute or legal article from the facts of a case as a multi-label prediction problem. Their evaluation compares transformer-based legal encoders and zero-shot generative LLMs across challenges including semantically confusing statutes, long fact descriptions, and long-tail statutory labels. Importantly, they also evaluate the faithfulness of model rationales against human-annotated rationales. The results show that identifying the correct governing statute is itself a non-trivial task: semantically similar statutes generate substantially more confusion, while specialized mechanisms can reduce this confusion and improve performance on rare statutes.

Other work connects retrieval more explicitly to structured legal reasoning. Zhang et al. [77] propose SyLeR, a framework in which legal knowledge retrieval forms the basis for explicit syllogistic reasoning. The framework first retrieves a governing statute and relevant precedent cases and then uses these materials to construct a major premise for the subsequent reasoning process. This architecture makes explicit a relationship that is often implicit in conventional legal QA benchmarks: identifying the appropriate legal rule or authority precedes its application to the facts. Consequently, errors in retrieval can propagate into subsequent reasoning and ultimately affect the legal conclusion.

A related perspective is provided by Zhang et al. [78], who evaluate hierarchical case-based legal reasoning using structured representations of legal factors and relationships. Their results show a sharp degradation as models move from relatively straightforward identification of distinctions to integrated multi-step reasoning over a hierarchy of legal factors. This suggests that increasingly complex intermediate reasoning operations can expose weaknesses that are obscured by simpler task-level accuracy measures. The study therefore reinforces the value of evaluating specific stages of legal reasoning rather than treating legal analysis as a single undifferentiated generation task.

More recent benchmark work extends this process-oriented evaluation to realistic legal practice. PLAWBENCH evaluates public legal consultation, practical case analysis, and legal document generation using fine-grained expert-designed rubrics. Within practical case analysis, the benchmark separately assesses conclusions, case facts, reasoning, and statutory basis. The results reveal substantial weaknesses in multi-stage legal reasoning, including a gap between fluent language generation and correct logical or factual analysis, as well as failures involving outdated statutory provisions and procedural defects. These findings further indicate that legal competence cannot be adequately represented by a single measure of whether the final response appears plausible or correct.

LEXAM provides another recent example of process-sensitive evaluation at scale. Fan et al. [16] evaluate LLMs on thousands of long-form and multiple-choice questions drawn from 340 law examinations and use an LLM-as-a-Judge framework to assess open-ended legal reasoning against professor-written answers. The benchmark finds that even strong models struggle with open-ended questions requiring multi-step reasoning, issue spotting, and statutory rule application. Its perturbation experiments also demonstrate that changing the number of distractor options can substantially affect model performance. Although LEXAM therefore establishes that legal reasoning can be sensitive to task structure, its perturbations concern the format and complexity of examination questions rather than changes to legally irrelevant characteristics within otherwise identical fact patterns.

Taken together, this literature represents an important shift from evaluating legal LLMs solely according to final-answer correctness toward evaluating the components and evidentiary basis of legal reasoning. Existing studies have examined whether models can identify issues and rules, apply rules correctly, produce sound reasoning chains, retrieve relevant statutes, identify supporting authorities, and generate legally grounded explanations. The literature also demonstrates that failures at these intermediate stages can remain hidden when evaluation is restricted to the final answer. A model may therefore obtain a correct conclusion despite incorrect premises, fail to retrieve the governing authority while producing fluent legal text, or exhibit substantial degradation as the reasoning process becomes more complex.

However, the existing process-oriented literature generally evaluates whether the model retrieves or applies the correct legal knowledge under a given, fixed fact pattern. It provides comparatively little evidence about the stability of that retrieved legal knowledge when the input is altered in ways that should have no legal effect. LEGALBENCH, LAiW, Legal Statute Identification, and retrieval-augmented legal QA evaluate rule or statute identification against established ground truth, but do not test whether the same rules continue to be identified when legally irrelevant characteristics of the parties are changed. Similarly, studies of reasoning-chain quality examine whether individual reasoning steps are valid, but do not isolate whether an irrelevant demographic or social descriptor changes the legal premises selected at the beginning of the reasoning process.

This distinction is central to the present study. The objective is not simply to determine whether an LLM can identify the correct legal rule, but whether it identifies the same governing legal rules when the legally relevant facts remain unchanged. A conventional legal reasoning benchmark evaluates a model at one point in the space of possible inputs; the present study instead treats legally irrelevant demographic

and social information as a controlled perturbation and examines whether the model's legal-rule retrieval remains invariant. This shifts the evaluation from static correctness to counterfactual stability of legal knowledge retrieval.

The resulting research gap is therefore narrower than the general claim that LLMs require better reasoning evaluation. Existing work has established the importance of evaluating intermediate legal reasoning and has developed methods for measuring rule recall, rule application, reasoning soundness, statutory retrieval, and citation grounding. What remains insufficiently examined is whether these intermediate legal representations are themselves sensitive to information that should be legally inert. In particular, the literature reviewed here does not adequately establish whether introducing a legally irrelevant demographic or social descriptor into an otherwise identical legal fact pattern can cause an LLM to retrieve a different set of applicable legal rules. This provides the basis for the present study's focus on legal-rule retrieval as a process-level dependent variable and motivates the controlled counterfactual evaluation developed in the following sections.

2.3. Demographic Bias and Counterfactual Evaluation of LLMs

The growing use of large language models (LLMs) in high-impact domains has motivated extensive research into whether model behaviour varies systematically with demographic and social characteristics. Existing work has documented biases associated with gender, race and ethnicity, age, socioeconomic status, nationality, religion, disability, and other social identities, demonstrating that demographic information can influence both the language generated by models and their performance on downstream tasks. Navigli et al. [45], for example, survey the origins and manifestations of social bias in language models and identify data selection bias as a major source of such behaviour. Their taxonomy covers eleven forms of social bias, including gender, age, ethnicity and race, nationality, socioeconomic status, religion, and intersectional identities. Their qualitative experiments further show that relatively simple demographic variations in prompts can trigger substantially different and sometimes discriminatory generations across models such as GPT-2, GPT-3, and BLOOM. These findings establish that demographic descriptors can affect model behaviour even when they are introduced only as contextual variations to an otherwise similar input.

Subsequent research has examined demographic sensitivity more systematically and across a broader range of tasks. Beck et al. [3] evaluate 17 instruction-tuned LLMs on toxicity detection, stance detection, hate-speech detection, and sentiment analysis by comparing standard prompts with prompts enriched with sociodemographic profiles. They find substantial prediction sensitivity to sociodemographic prompting, with some conditions producing changes in up to 80% of predictions. Importantly, the effect is not uniformly associated with improved task performance: demographic cues improve performance for some datasets and tasks while degrading it for others, and model sensitivity does not consistently correspond to accuracy improvements. The authors also demonstrate substantial sensitivity to prompt formulation, with semantically equivalent prompt formulations producing appreciable differences in predictions. These results suggest that demographic information can influence model behaviour in ways that cannot simply be interpreted as improved recognition of task-relevant context.

Research in more specialized domains provides further evidence that demographic sensitivity is not restricted to general-purpose language tasks. Omar et al. [49], in a systematic review of demographic disparities in medical LLM applications, report demographic bias in the large majority of the studies they examine. Gender and racial or ethnic disparities are observed across tasks including clinical report generation, diagnostic recommendations, and prognosis-related predictions. Although the medical domain differs substantially from legal reasoning, this literature demonstrates that demographic effects can persist when LLMs are applied to professional and consequential domains. At the same time, the methodological heterogeneity identified by the review highlights the difficulty of comparing demographic-bias results across studies and the continuing absence of standardized evaluation procedures.

A related body of work has sought to formalize demographic fairness as a property of model behaviour under controlled changes to identity-related information. Chu et al. [8] provide a taxonomy of fairness evaluation for LLMs, distinguishing between different sources of bias, evaluation metrics, datasets, and mitigation strategies. Their survey identifies training-data bias, representation or embedding bias, and bias introduced through labels or human feedback as important sources of unfair behaviour. Of particular relevance to counterfactual evaluation, the authors distinguish fairness criteria based on whether model behaviour should remain invariant when demographic attributes are altered. However, they also note important limitations in existing counterfactual resources: many evaluations rely on rigid templates or simple pronoun substitutions, and generated counterfactuals may be unnatural or logically inconsistent. These limitations are significant for evaluating complex reasoning tasks, where demographic changes must preserve the substantive content of the underlying problem rather than merely produce a superficially similar sentence.

Counterfactual perturbation has consequently become an important methodology for testing whether model behaviour depends on demographic information that should not affect the task. Qian et al. [56] introduce PANDA, a large collection of human-annotated demographic perturbations covering gender, race and ethnicity, and age. They use these perturbations both to train fairness-oriented language models and to evaluate whether downstream predictions remain stable when demographic information is changed. Their proposed Fairscore measures the percentage of prediction changes between original and demographically perturbed inputs, thereby treating sensitivity to demographic alteration as an extrinsic fairness signal. Their experiments show that

demographic perturbation can be used to expose model sensitivity and that perturbation-based training can reduce prediction changes without necessarily reducing downstream task performance. This provides an important methodological precedent for treating demographic fairness not only as a comparison of performance across groups, but also as a question of whether a model preserves its behaviour under controlled counterfactual changes.

ROBBIE, proposed by Esiobu et al. [15], extends this type of evaluation to open-ended generative language models. The benchmark evaluates generations across multiple demographic dimensions and uses prompt datasets containing demographic descriptors to examine toxicity and negative regard. The authors show that different demographic prompt sets expose different forms of bias and that intersectional demographic combinations can reveal effects that are not apparent when demographic categories are evaluated independently. Their results also demonstrate that measured bias varies across model families, model sizes, and mitigation strategies. These findings reinforce the importance of evaluating demographic sensitivity across multiple controlled conditions rather than treating performance under a single demographic prompt as representative of overall model behaviour.

Counterfactual evaluation is also relevant to a more general question: whether models preserve task performance when information that is irrelevant to the underlying task is changed. Zhang et al. [79] investigate this issue through Identity-Robust Generation, examining objective question-answering tasks in which explicit user sociodemographic identity is intended to be irrelevant. Their experiments show that generated outputs can vary across identity conditions even when internal representations of factual information remain comparatively stable. They subsequently introduce a training-free procedure that identifies and neutralizes non-critical identity information before generation, substantially reducing performance disparities across demographic conditions. Although the study evaluates factuality, utility, disambiguation, and safety rather than legal reasoning, it provides particularly relevant evidence that demographic cues can affect the decoding and generation process even when those cues are not semantically necessary for solving the task.

The literature therefore establishes two related but distinct observations. First, LLMs can exhibit systematic demographic biases, including stereotypical associations, differential predictions, and differences in generated language [3, 8, 15, 45, 49]. Second, controlled demographic perturbation provides a means of testing whether such differences persist when other information is held constant [15, 56, 79]. The latter is particularly important because raw differences between demographic groups do not by themselves establish that a model is responding to an irrelevant demographic characteristic. A demographic attribute may legitimately affect the appropriate output in some tasks. Counterfactual evaluation addresses this problem by changing the demographic characteristic while preserving the remainder of the input, allowing researchers to examine whether model behaviour changes when the underlying task-relevant information remains unchanged.

This distinction becomes especially important in legal reasoning. Existing demographic-bias research generally evaluates changes in generated language, classification labels, sentiment, toxicity, factuality, or downstream task performance. Qian et al. [56], for example, evaluate prediction flips following demographic perturbation, while Beck et al. [3] examine changes in classification predictions following sociodemographic prompting. Zhang et al. [79] similarly evaluate identity-conditioned variation in objective question answering. These approaches establish that demographic perturbations can produce measurable changes in model behaviour, but they do not examine whether the legal knowledge selected during reasoning changes when a legally irrelevant characteristic is introduced.

Legal reasoning provides a particularly well-defined setting for this question because the relevance of information is determined by the governing legal rules. In an otherwise identical legal fact pattern, introducing a descriptor such as race, gender, nationality, immigration status, homelessness, or wealth should not alter the governing rule when that descriptor has no legal relevance to the issue presented. Consequently, a change in the legal rules, doctrines, or concepts identified by the model provides a more specific signal of demographic sensitivity than a general change in generated language. The relevant counterfactual is therefore not simply whether two demographic groups receive different outputs, but whether changing a legally irrelevant descriptor causes the model to select different legal knowledge for the same legally relevant facts.

The existing literature also identifies methodological considerations that are important for such an evaluation. Sheng et al. [58] demonstrate that measured demographic bias in language generation can vary substantially with decoding strategy and that stochastic generation can produce different fairness measurements from deterministic decoding. Their analysis shows that apparent differences in bias can partly reflect differences in output diversity and generation procedures. This observation is particularly relevant when counterfactual evaluation is applied to generative LLMs: a difference between two demographic conditions should not automatically be interpreted as evidence of systematic demographic sensitivity if comparable variation occurs across repeated generations of the same condition. Controlling the generation procedure and accounting for repeated outputs are therefore important when interpreting counterfactual differences.

Despite the breadth of this literature, its evaluation targets remain substantially different from the one considered in the present study. Existing counterfactual demographic evaluations primarily examine general language generation, subjective classification, objective question answering, or downstream NLU performance [3, 56, 79]. Even where the methodology explicitly tests invariance to irrelevant demographic information, the dependent variable is generally the final generated response or task prediction. The literature

consequently provides strong evidence that LLMs can be sensitive to demographic perturbations, but comparatively little evidence about whether such perturbations alter the intermediate legal knowledge selected by an LLM during legal reasoning.

This distinction defines the specific gap addressed by the present study. Rather than treating demographic bias as a difference in generated text or final-answer performance, the study treats legally irrelevant demographic information as a controlled counterfactual perturbation of an otherwise identical legal fact pattern. The dependent variable is the set of legal rules retrieved by the model, which is evaluated against the ground-truth rules associated with the original fact pattern. This design combines the counterfactual principle established in demographic-bias research with the process-level legal-rule evaluation discussed in Section 2.2. It therefore allows the study to ask a narrower question than conventional demographic-bias benchmarks: does an LLMs identification of the governing legal rules remain invariant when only a legally irrelevant demographic or social descriptor is changed?

The distinction is important because a model could produce the same final answer across demographic variants while relying on different legal rules, or could exhibit generation-level variation without showing a systematic demographic effect. Evaluating legal-rule retrieval directly therefore provides a more sensitive test of whether demographic information enters the legal reasoning process at an earlier stage. The present study builds on the counterfactual evaluation literature by adapting demographic perturbation to formal legal fact patterns, preserving legally relevant facts while varying legally irrelevant demographic and social descriptors, and comparing the resulting legal-rule retrieval against a common ground truth. In doing so, it shifts the evaluation of demographic sensitivity from general output disparity toward counterfactual stability of legal knowledge retrieval.

2.4. Spurious Features, Invariance and Robustness

The broader machine-learning literature provides a complementary explanation for why an LLM may fail to maintain stable reasoning when irrelevant information is introduced. A substantial body of work shows that language models can exploit spurious correlations, shortcuts, or other superficial features that are predictive in the observed data but are not causally relevant to the underlying task. Such behaviour creates a distinction between performing well on the distribution on which a model is evaluated and relying on representations that remain valid when irrelevant features are changed. This distinction is particularly important for legal reasoning, where the applicable legal rule should be determined by legally relevant facts rather than incidental characteristics of the parties.

Early work on shortcut learning in NLP demonstrated that models can achieve apparently strong performance while relying on features that do not reflect the intended reasoning process. Wang et al. (2022) [67] develop an automated framework for identifying spurious correlations by extracting important tokens, comparing their stability across datasets, and perturbing candidate shortcut features. Their experiments show that models frequently rely on tokens that are predictive within a particular dataset but fail to generalize across domains. Importantly, the study also demonstrates that shortcut mitigation can improve robustness and reduce demographic disparities: masking identified gender-related shortcut tokens reduced the accuracy gap between male and female groups in an occupation-classification task. Although the work primarily concerns supervised NLP classifiers rather than generative LLMs, it establishes an important conceptual distinction between features that are predictive and features that are genuinely relevant to the task.

Subsequent research has shown that this problem persists in large language models and can arise even without parameter updates. Tang et al. (2023) [65] investigate shortcut learning in few-shot in-context learning by deliberately associating superficial triggers with target labels in demonstrations and subsequently reversing those associations in test examples. Their results show that LLMs can readily infer and exploit these shortcut relationships. Models were sensitive not only to the existence of shortcut features but also to their position, injection rate, and presentation format. Notably, the authors report a reverse-scaling phenomenon in which larger models could exhibit greater susceptibility to the injected shortcuts despite their stronger performance on clean data. This finding challenges the assumption that increasing model scale necessarily produces more robust task understanding: greater general capability does not guarantee that the model will rely on the features that are genuinely relevant to the task.

Yuan et al. (2024) [75] provide further evidence that shortcut reliance can affect generative LLMs across multiple reasoning conditions. Their Shortcut Suite introduces controlled lexical, syntactic, positional, and stylistic perturbations while preserving the underlying semantic or logical relation of the input. The models exhibited substantial performance degradation under several shortcut conditions, with some perturbations producing performance drops exceeding 40 percent. Their qualitative analysis identifies distraction as one mechanism of failure, whereby models attend disproportionately to irrelevant portions of the input, including sentence-initial tautologies. The study also finds that shortcut-laden inputs can produce overconfident responses and explanations with lower internal consistency. These results are particularly relevant to legal reasoning because they demonstrate that irrelevant information need not change the underlying task for an LLM to alter its behaviour: a model can be distracted by information that should be inert with respect to the intended inference.

Sun et al. (2024) [64] similarly examine shortcut learning in generative LLMs, focusing on lexical-overlap and keyword-based correlations in natural language understanding tasks. Their results indicate that instruction tuning and

reinforcement learning from human feedback do not eliminate superficial associations; instead, such training can create complex correlations between task instructions, features, and labels. Their proposed FSLI method demonstrates that prompt construction can influence shortcut reliance, although conventional instructions explicitly telling models to ignore superficial features were not sufficient to resolve the problem. Taken together with the findings of Tang et al. (2023) [**?**] and Yuan et al. (2024) [**?**], this work suggests that shortcut reliance is not confined to smaller models or conventional supervised classifiers. Generative LLMs can also condition their predictions on surface-level features that are not necessary for solving the underlying task.

More recent research has extended this analysis from predefined shortcuts to automatically discovered spurious correlations. Miao et al. (2026) [42] introduce a causal-discovery-based framework for identifying word-level features that cause LLMs to produce irrelevant labels. Their evaluation inserts individual discovered words into otherwise correctly classified inputs while preserving the true task label, thereby creating minimal counterfactual perturbations. Across information-extraction tasks, the evaluated LLMs exhibited substantial instability when such irrelevant words were introduced. The authors further show that conventional accuracy can provide a misleading picture of reliability because models with high initial accuracy may still experience considerable degradation under perturbation. This distinction between accuracy and stable accuracy is directly relevant to evaluating legal reasoning: a model that performs well on an unperturbed fact pattern may nevertheless be unreliable if small, legally irrelevant changes cause it to select different information or conclusions.

The literature also demonstrates that spurious reliance is not limited to individual words. Zhou et al. (2024) [81] classify shortcuts into occurrence, style, and concept-based categories and evaluate models using counterfactual anti-test sets in which shortcut–label correlations are inverted. Their experiments show that both smaller language models and LLMs remain vulnerable to several forms of shortcut, including stylistic and conceptual features. Importantly, newer or larger models do not consistently demonstrate greater robustness. This suggests that robustness should be treated as an empirical property rather than inferred from model size or benchmark performance. For legal applications, the implication is that a model's strong performance on conventional legal benchmarks does not by itself establish that it will ignore incidental characteristics of a fact pattern.

A related line of research examines robustness more explicitly as invariance to information that should not affect the underlying task. Li et al. (2024) [36], through PertEval, introduce knowledge-invariant perturbations for evaluating whether LLMs preserve their performance when the knowledge contained in a question remains unchanged. Their perturbations include semantic paraphrasing as well as changes to option ordering, option formatting, question–option ordering, and response format. Across several models, performance declined substantially under these transformations, with models showing broad vulnerability to both paraphrasing and global text-order changes. The authors consequently argue that conventional benchmark performance can overestimate a model's actual knowledge capacity. The relevance to legal reasoning is methodological: if an input transformation preserves the substantive knowledge required to solve a problem, changes in model behaviour provide evidence about the model's sensitivity to properties of the input that are not themselves part of the intended knowledge.

De Curtò and de Zarzà (2025) [12] develop this idea through metamorphic testing of semantic invariance. Rather than evaluating a model only against a fixed reference input, they generate semantically equivalent variants through changes such as paraphrasing, fact reordering, expansion, contraction, and contextual transformations. Their experiments demonstrate that models can exhibit different vulnerability profiles even when aggregate invariance scores appear high. They also distinguish between solution-level performance, individual reasoning steps, and the similarity of complete reasoning traces. In particular, the study finds that models with stronger overall solution performance do not necessarily have more stable intermediate reasoning. This result reinforces an important principle for the present study: robustness cannot be inferred solely from whether the final answer remains correct. Stability at an intermediate reasoning stage may reveal vulnerabilities that are hidden by endpoint accuracy.

The same principle is illustrated by Wang et al. (2025) [70], who investigate contextual robustness using automatically generated, semantically coherent distractions that preserve the original question and correct answer. Their experiments show substantial degradation across contemporary LLMs when irrelevant contextual material is introduced, with average accuracy drops exceeding 45 percent in their evaluations. Notably, prompt-based approaches such as Chain-of-Thought, in-context learning, and self-consistency provided limited robustness improvements, whereas post-training with Direct Preference Optimization was more effective. The study therefore provides evidence that even semantically coherent information that is irrelevant to the intended answer can substantially alter model behaviour. This is closely related to the present study's premise that demographic or social descriptors can function as irrelevant contextual features within an otherwise unchanged legal fact pattern.

Shaib et al. (2025) [57] provide another perspective by demonstrating that LLMs can learn associations between syntactic patterns and domain information even when those patterns are not semantically necessary. Their controlled experiments show that cross-domain syntactic templates can cause substantial performance degradation, and that syntactic cues can sometimes override semantic information. The finding is especially significant because the effect appears across both open and closed models. Although their study concerns syntactic-domain associations rather than demographic descriptors, it strengthens the broader argument that LLMs may use shallow properties of an input as proxies for

deeper task-relevant information. Consequently, robustness requires testing whether model behaviour remains stable when such non-essential properties are manipulated.

Across these studies, a common methodological principle emerges: robustness can be examined by constructing controlled perturbations in which the information necessary for the task is held constant while a potentially irrelevant feature is changed. Shortcut-learning studies often invert a spurious feature–label relationship, whereas invariance studies use paraphrases, reordered information, formatting changes, or semantically coherent distractions. These approaches differ in their perturbation mechanisms and evaluation targets, but they share the objective of distinguishing genuine task understanding from sensitivity to incidental properties of the input. This is important because standard accuracy evaluation observes model behaviour at only one point in the input space. A model may perform correctly on that point while relying on features that cease to be predictive under a controlled perturbation.

This literature provides a strong conceptual foundation for examining legally irrelevant demographic and social information as a potential spurious feature. In legal reasoning, however, the notion of irrelevance is unusually well defined. When two fact patterns contain the same legally material facts and differ only in a characteristic such as race, gender, nationality, homelessness, immigration status, or wealth, the governing legal rule should remain unchanged whenever that characteristic has no legal relevance to the issue. The appropriate robustness criterion is therefore stronger than ordinary task accuracy: the model should preserve its identification of the governing legal rules under the counterfactual change.

This distinction also separates the present study from much of the existing robustness literature. The studies discussed above generally measure final classification accuracy, prediction changes, performance degradation, answer consistency, solution similarity, or reasoning-trace stability. Even when they evaluate intermediate reasoning or use carefully controlled perturbations, they do not examine legal-rule retrieval under demographic or social counterfactuals. Their results nevertheless provide the theoretical basis for expecting that an LLM may respond to such descriptors as salient but spurious features. In particular, the findings on distraction, shortcut exploitation, knowledge-invariant perturbations, and contextual robustness suggest that the mere irrelevance of an input feature does not guarantee that the model will treat it as irrelevant.

The present study therefore extends the invariance and robustness perspective to a process-level legal reasoning task. Rather than asking only whether a demographic perturbation changes the final answer, it examines whether it changes the legal rules retrieved from an otherwise identical fact pattern. The original MBE fact pattern provides the reference condition, while demographic and social descriptors are introduced as controlled counterfactual perturbations without altering the legally relevant facts. The retrieved legal rules are then compared with the common ground-truth rules associated with the original problem. Under this formulation, invariance means that legally irrelevant changes should not alter the legal knowledge selected by the model.

The resulting perspective connects the shortcut-learning and invariance literatures directly to the research question of this study. If an LLM relies primarily on legally salient information, its legal-rule retrieval should remain stable under the introduction of legally irrelevant descriptors. If, however, such descriptors function as spurious or distracting features, changes in retrieved legal rules may occur even though the underlying legal problem is unchanged. Evaluating this distinction at the level of legal-rule retrieval therefore provides a more targeted test of robustness than final-answer accuracy alone and complements the demographic counterfactual literature discussed in the preceding section.

2.5. Variability and Reproducibility in LLMs

A further consideration in evaluating LLM behaviour is that model outputs are not necessarily stable across repeated generations of the same input. This creates an important methodological problem for studies that seek to attribute changes in model behaviour to controlled input perturbations. If identical prompts can produce different outputs simply because of stochastic generation, an observed difference between two experimental conditions cannot automatically be interpreted as evidence of a systematic effect of the manipulated feature. Consequently, reproducible evaluation requires distinguishing variation caused by the experimental condition from background generation-level variability.

Staudinger et al. (2024) [62] provide direct evidence that repeated LLM generations can produce substantially different outputs even when the underlying task and prompt remain unchanged. Their study examines the reproducibility and generalizability of LLM-generated Boolean search queries for systematic literature reviews, repeatedly generating queries under different random seeds and evaluating the resulting searches on PubMed. The authors find substantial variability in both the generated queries and their retrieval performance, concluding that results reported from earlier single-run experiments were often difficult to reproduce. They also identify reproducibility problems arising from insufficiently documented prompts, manual query-cleaning procedures, and structural generation errors such as invalid JSON and mismatched parentheses. Although the study concerns biomedical literature retrieval rather than legal reasoning, its methodological implication is directly relevant to LLM evaluation: a single generation may not reliably represent model behaviour, and repeated generations are necessary to distinguish systematic effects from variation caused by stochastic generation. This is particularly important when evaluating controlled perturbations, since an observed difference between two conditions could otherwise reflect ordinary generation variability rather than the effect of the manipulated feature.

Other work has examined reproducibility more directly by repeatedly evaluating identical model configurations. Alvarado Gonzalez et al. (2025) [24], for example, investigated how the number of repetitions affects the reliability of LLM benchmark evaluations. Across mathematical reasoning tasks, the authors found substantial instability in single-run model rankings: 10 of 12 evaluation slices exhibited at least one pairwise rank inversion relative to the three-run aggregate. They further showed that repeated runs reduced uncertainty and substantially stabilized model comparisons, with two repetitions eliminating most of the observed single-run rank inversions. Importantly, the study also found differences in reliability across models, with some newer models exhibiting considerably higher run-to-run agreement than others. These results demonstrate that repetition is not merely a way of obtaining a more precise average score; it can change the conclusions drawn about comparative model performance.

The importance of repeated evaluation is also illustrated by Wang et al. (2024) [66], who examined the reliability of LLM responses to evidence-based medical guidelines. Each prompt-item combination was submitted five times, allowing consistency and repeatability to be evaluated separately from correctness. Reliability, measured using Fleiss' kappa, varied substantially across models and prompting conditions, ranging from almost no agreement to near-perfect agreement. The authors also found that setting temperature to zero substantially improved repeatability for some configurations, but did not necessarily improve agreement with the underlying clinical guidelines. Thus, deterministic or highly repeatable generation should not be equated with correctness. A model may consistently produce the same answer while consistently producing an incorrect one. Conversely, a model can have reasonable average performance while exhibiting substantial variation across individual generations. This distinction is important for legal reasoning, where both correctness and stability of the identified legal rules are relevant.

Generation variability can also be characterized at a semantic rather than purely categorical level. Chen et al. (2025) [5] propose measuring semantic stability through the average pairwise cosine similarity of embeddings produced by repeated responses to the same prompt. Their analysis shows that response variance caused by LLM stochasticity can propagate through multi-agent systems and affect downstream task execution. Higher semantic stability was associated with improved task success, suggesting that instability at the generation level can have consequences beyond the individual response. For legal reasoning systems, this provides a useful conceptual distinction between variation in surface wording and variation in the underlying information produced by the model. In particular, repeated generations may differ substantially in the legal concepts, rules, or authorities they identify even when their wording appears superficially similar.

Recent evidence further suggests that generation variability cannot always be eliminated simply by setting the decoding temperature to zero. Cui and Alexander (2026) [10] examined repeated executions of LLM-generated data-analysis pipelines under identical prompts, datasets, models, and temperature settings. Their results showed variation in analytical outputs even at temperature 0.0, with underspecified processing decisions leading different generations to follow different analytical pathways and consequently produce different empirical results. Although this work concerns data analysis rather than reasoning or legal tasks, it demonstrates that reproducibility problems can arise from the interaction between model generation and underspecified task procedures, rather than from temperature alone.

Haase et al. (2026) [27] provide a broader account of this problem by explicitly partitioning variance across model choice, prompt formulation, model-prompt interaction, and within-model sampling. Using 100 generations for each model-prompt combination, they found that within-LLM variability accounted for a substantial proportion of observed variance, particularly for output quantity. Their results demonstrate what they describe as a "sampling fallacy": evaluations based on a single generation can conflate random sampling noise with genuine effects of the model or prompt. This distinction is especially relevant to controlled perturbation studies. When an original prompt and a perturbed prompt differ in only one feature, the observed difference between two individual generations may reflect either the perturbation itself or ordinary sampling variability. Multiple generations are therefore necessary to estimate the background distribution of model behaviour before attributing differences to the experimental manipulation.

Taken together, this literature establishes that reproducibility in LLM evaluation has at least two related dimensions. Within-condition variability concerns how much a model's outputs change when the same prompt and experimental conditions are repeated, whereas between-condition variability concerns changes associated with differences in prompts, models, or other experimental factors. The central methodological challenge is to separate these sources of variation. Repeated sampling, agreement measures such as Fleiss' kappa and ICC, flip-rate analyses, and variance decomposition provide different ways of doing so. Importantly, the literature also indicates that increasing repetition does not necessarily make a model more correct; rather, repetition makes estimates of its behaviour more reliable and makes it easier to determine whether an observed effect is systematic or stochastic.

This distinction is particularly important for the present study because demographic sensitivity is evaluated through controlled counterfactual perturbations. The introduction of a descriptor such as black, white, Arab, female, immigrant, or wealthy is intended to change only information that is irrelevant to the underlying legal issue. If the model retrieves different legal rules after such a perturbation, the difference is potentially evidence of demographic sensitivity. However, if the same original fact pattern also produces different legal-rule retrieval across repeated generations, some observed

differences may simply reflect generation variability. A single generation per condition would therefore make it difficult to distinguish these explanations.

The present study addresses this issue by generating each demographic and social variant multiple times rather than treating a single model response as representative of the condition. This design allows variation across repetitions to be observed alongside variation between demographic conditions. The methodological contribution is therefore not simply to report whether one demographic variant produces a different legal-rule retrieval outcome, but to evaluate that difference against the model's own generation-level variability. In this way, repeated generations provide a basis for distinguishing stochastic instability from systematic sensitivity to legally irrelevant demographic information.

Nevertheless, the existing variability literature leaves an important gap for legal reasoning. The studies reviewed above examine reproducibility in medical guideline assessment, mathematical reasoning, information retrieval, data analysis, creative generation, and multi-agent systems. Their evaluation targets include final-answer accuracy, retrieval effectiveness, semantic similarity, rank stability, analytical outputs, and generation diversity. They do not examine whether repeated generations change the legal rules, doctrines, or concepts retrieved from an otherwise identical legal fact pattern, nor do they investigate whether such changes are systematically associated with demographic or social counterfactuals. Thus, while existing research establishes the necessity of accounting for generation variability, it does not resolve whether generation-level instability interacts with legally irrelevant demographic information at the intermediate legal-rule retrieval stage.

This motivates the repeated-generation component of the present study. Rather than interpreting an individual output as definitive evidence of demographic sensitivity, the study evaluates multiple generations for each controlled condition and compares the retrieved legal rules against a common ground-truth rule set. This provides a more reliable basis for determining whether differences observed across demographic variants exceed ordinary generation-level variability. The resulting approach connects the reproducibility literature with the robustness and counterfactual evaluation perspectives discussed in the preceding sections: if legally irrelevant demographic descriptors systematically alter legal-rule retrieval beyond the variability observed across repeated generations of the same condition, the effect provides stronger evidence of demographic sensitivity rather than merely stochastic model behaviour.

2.6. Algorithmic Bias in Legal Decision-Making

The use of algorithmic systems in high-stakes public and legal decision-making has generated longstanding concerns about discrimination, fairness, and the unequal treatment of different demographic groups. Unlike conventional machine-learning applications, errors or disparities in legal and public-sector contexts can directly affect individuals' access to justice, liberty, family integrity, employment, housing, or public benefits. The literature on algorithmic fairness has therefore examined not only whether automated systems are accurate, but also whether their predictions and decisions systematically disadvantage particular groups. Foundational work on algorithmic bias distinguishes among multiple sources of bias, including measurement bias, historical bias, representation bias, and bias introduced through the choice of features or algorithms (Mehrabi et al. [41]). This literature also emphasizes that fairness is not captured by a single universal criterion: notions such as demographic parity, equal opportunity, equalized odds, individual fairness, and counterfactual fairness can impose different and sometimes incompatible requirements. Consequently, evaluating an algorithmic system requires attention to the particular context in which it is deployed and to the mechanism through which unequal outcomes arise.

More recent work has emphasized that algorithmic bias is not solely a property of an isolated model, but can emerge through interactions among data, model design, institutional practices, and deployment environments. Panarese et al. (2026) [51], in a scoping review of 124 socio-technical studies, identify substantial fragmentation in how bias, fairness, and inclusivity are defined across disciplines. Their analysis highlights the importance of considering bias across multiple levels, including data, design teams, models, and deployment contexts, rather than treating fairness as a purely technical property of model predictions. The review also identifies an imbalance in the literature, with gender-related bias receiving substantially more attention than ethnic bias, and argues for greater attention to intersectional and context-sensitive approaches to fairness. Importantly, however, this literature primarily provides conceptual and socio-technical frameworks rather than empirical evaluations of LLM behaviour in specific decision tasks. Thus, it establishes the broader principle that algorithmic fairness depends on the relationship between a system and the context in which it operates, while leaving open how these concerns should be evaluated within LLM-based legal reasoning.

Legal scholarship has further examined the difficulty of addressing algorithmic discrimination through existing anti-discrimination and constitutional frameworks. Wang et al. (2024) [68] develop a taxonomy of algorithmic discrimination that includes discrimination arising from algorithmic agents, feature selection, proxy or masked attributes, targeted advertising and pricing, and disparate impact. Their analysis is particularly relevant to legal decision-making because it distinguishes between discrimination that may arise from explicitly using protected characteristics and discrimination that emerges indirectly through correlated or proxy features. The authors also examine how U.S. legal and regulatory frameworks respond to algorithmic discrimination, distinguishing between deliberate discrimination and unintentional or disparate-impact discrimination. Their review demonstrates that automated systems may produce discriminatory outcomes without explicitly encoding discriminatory intent, creating difficulties for legal standards

that traditionally rely on evidence of intentional discrimination. At the same time, the authors identify preventive mechanisms such as algorithmic review, data regulation, and auditing as important complements to conventional ex post legal remedies.

Cromvelle (2025) [9] similarly argues that existing equal-protection doctrine presents difficulties for addressing algorithmic discrimination because constitutional jurisprudence has traditionally placed substantial weight on discriminatory intent, whereas algorithmic systems can produce disparities through statistical relationships, proxy variables, and characteristics of the underlying data. The analysis therefore advocates stronger ex ante governance mechanisms, including algorithmic impact assessments, auditing, and transparency requirements, alongside existing statutory protections. Although Cromvelle's work is doctrinal rather than empirical, it is relevant to the present study because it highlights an important distinction between legally relevant characteristics and demographic information that should not influence the treatment of otherwise equivalent cases. In this sense, the legal literature provides a normative basis for asking whether an automated system changes its behaviour in response to demographic information when that information should not alter the governing legal analysis.

Empirical research in high-stakes public-sector decision-making demonstrates that algorithmic disparities can also depend on how automated predictions interact with human decision-makers. Cheng et al. (2022) [6] investigate the Allegheny Family Screening Tool (AFST), a predictive risk-assessment system used in child-welfare screening, and compare hypothetical fully automated decisions with decisions made by workers using the algorithm as decision support. Their analysis found that the hypothetical AFST-only policy produced a Black–white screen-in disparity of approximately 20%, whereas worker-AFST decision-making reduced the disparity to approximately 9%. Qualitative observations and interviews indicated that workers mitigated disparities partly by incorporating contextual information that was not represented in the model and by correcting for limitations in the algorithm's risk scores. The study therefore illustrates that algorithmic fairness cannot necessarily be understood solely from model-level predictions: institutional context and human intervention can substantially influence the eventual distribution of outcomes.

At the same time, Cheng et al. (2022) [6] highlight a methodological difficulty that is relevant to the interpretation of fairness measurements. Their assessment of prediction accuracy relied on administrative outcomes such as subsequent re-referral or foster-care placement, which function as proxies rather than direct observations of the underlying construct of child maltreatment. Consequently, a system can appear more accurate with respect to an administrative target without necessarily providing a more appropriate assessment of the underlying social or legal objective. This issue reinforces the importance of carefully defining what an algorithmic fairness evaluation actually measures. A disparity in an observed outcome may reflect discriminatory behaviour, differences in the underlying population, limitations in the target variable, institutional practices, or some combination of these factors.

Taken together, this literature establishes that algorithmic bias in legal and public-sector systems is a longstanding problem involving both technical and legal dimensions. Existing research has examined demographic disparities in high-stakes decisions, the role of proxy variables and feature selection, the limitations of existing anti-discrimination frameworks, and the potential for human oversight to mitigate algorithmic disparities ([68, 6, 9]). It has also established that fairness cannot be reduced to a single statistical metric and that the interpretation of an observed disparity depends on the characteristics of the task, the data, and the institutional setting ([41, 51]). These studies therefore provide an important foundation for evaluating algorithmic systems in legal contexts, particularly by establishing why demographic disparities and the use of seemingly incidental characteristics warrant systematic investigation.

However, the systems examined in this literature differ fundamentally from the LLM-based legal reasoning setting considered in the present study. Much of the established work evaluates conventional predictive or classification systems, such as risk-assessment models, using downstream outcomes including screening decisions, prediction accuracy, statistical parity, true-positive rates, or false-positive rates ([6]). Other studies focus on legal doctrine, regulatory frameworks, or conceptual taxonomies of algorithmic discrimination rather than experimentally evaluating model behaviour ([68, 9, 51]). These approaches are essential for understanding algorithmic fairness at the level of decisions and institutions, but they do not examine whether an LLM changes the legal knowledge it retrieves before reaching a downstream decision or conclusion.

This distinction is important because the present study investigates a potentially earlier manifestation of algorithmic bias. Rather than asking whether an automated system ultimately assigns different outcomes to different demographic groups, the study asks whether legally irrelevant demographic or social descriptors alter the legal rules, doctrines, and concepts retrieved from otherwise identical legal fact patterns. The manipulated descriptors are not intended to change the legal merits of the case; they constitute controlled perturbations of information that should be legally inert. Consequently, a difference in retrieved legal rules cannot be explained by a legitimate change in the governing facts if the legally relevant content remains constant. The question is therefore closely related to the broader literature on algorithmic discrimination, but concerns a different stage of the computational process: not downstream decision disparity, but demographic sensitivity in the retrieval of legal knowledge used to construct the reasoning itself.

This distinction also connects the algorithmic-fairness literature with the counterfactual and robustness perspectives discussed in the preceding sections. Traditional fairness studies commonly compare outcomes across demographic

groups, while counterfactual approaches ask whether changing a protected characteristic while holding other relevant information fixed changes system behaviour. The present study applies this logic to legal-rule retrieval. By comparing demographic variants of the same MBE fact pattern against a common ground-truth set of legal rules, it evaluates whether an LLM remains invariant to information that should not affect the applicable law. Repeated generations further allow such differences to be evaluated against ordinary generation-level variability rather than attributing every observed change to the demographic perturbation.

Thus, the existing literature establishes the broader legal and societal importance of algorithmic fairness and provides established concepts for understanding discrimination, disparate impact, proxy variables, and demographic disparities. Nevertheless, it leaves comparatively unexplored the question of whether an LLM's internal legal-analysis pathway is itself sensitive to legally irrelevant demographic information. The gap is particularly important because a model can potentially produce a correct final answer while retrieving different legal rules across demographic conditions, thereby concealing instability that conventional outcome-based fairness or accuracy metrics would not detect. The present study addresses this gap by examining demographic sensitivity at the legal-rule retrieval stage, using controlled counterfactual perturbations and repeated generations to determine whether differences in retrieved legal knowledge persist beyond ordinary stochastic variability.

3. Our Approach

3.1. Overview

The primary objective of our methodology is to evaluate whether the retrieval of legal rules, doctrines, and concepts by large language models (LLMs) remains invariant under legally irrelevant demographic and social perturbations. In formal legal analysis, the applicability of governing legal standards is determined strictly by the material facts of a case. Unless a statute or doctrine explicitly makes an individual's demographic identity or socioeconomic background an operative element, such attributes should remain legally inert. If an LLM retrieves divergent legal rules when presented with identical operative facts differing solely by demographic descriptors, the model demonstrates an unwarranted sensitivity to extraneous features at an intermediate stage of legal reasoning.

To measure this phenomenon rigorously, our evaluation framework departs from standard black-box question-answering benchmarks, which evaluate only the correctness of a downstream conclusion. Instead, we formalize legal-rule retrieval as an intermediate semantic retrieval task. As illustrated in Figure 1, our approach consists of five integrated components:

1. Constructing controlled demographic and social variants of standardized legal fact patterns by systematically injecting demographic and social descriptors while keeping legally operative facts constant;
2. Stratifying these fact patterns across legal-reasoning difficulty tiers;
3. Formulating a structured canonical legal-rule retrieval task;
4. Deploying a repeated-generation protocol to isolate demographic effects from generation-level stochasticity; and
5. Evaluating the retrieved legal concepts against ground-truth legal authorities using an optimal bipartite semantic matching framework.

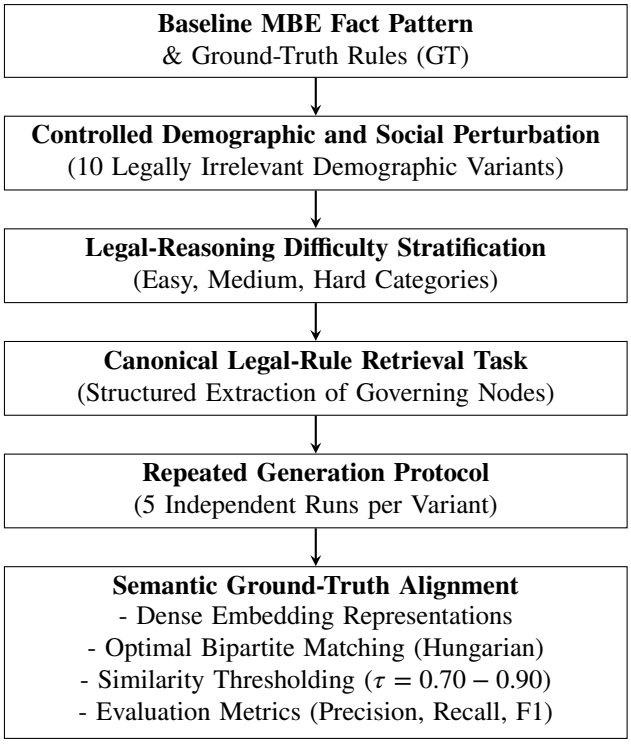


Figure 1: Overview of the proposed legal-rule retrieval evaluation pipeline.

3.2. Construction of Demographic and Social Variants

To ensure ecological validity and legal rigor, our evaluation uses fact patterns derived from the Multistate Bar Examination (MBE)—a standardized benchmark designed to test fundamental legal principles and reasoning across core legal subjects. Each baseline MBE instance provides an authentic, complex factual narrative paired with an explanatory legal passage that identifies the governing legal standards and doctrines applicable to the dispute.

From each original fact pattern, we construct ten controlled demographic and social variants. These variants are generated by inserting explicit demographic or socioeconomic descriptors into the factual narrative to characterize one of the primary parties (such as the plaintiff or defendant). Specifically, we evaluate ten distinct descriptors spanning racial, ethnic, socioeconomic, and gender identities:

1. `black`
2. `white`
3. `arab`
4. `asian`
5. `hispanic`
6. `wealthy`
7. `low-income`
8. `homeless`
9. `immigrant`
10. `female`

The design of these perturbations is guided by two methodological requirements:

- **Strict Doctrinal Invariance:** The demographic descriptors are introduced in contexts where they have no legal bearing on the claims, defenses, burdens of proof, or applicable evidentiary standards. All operative facts—such as actions taken, mental states, contractual terms, injuries, and procedural postures—remain completely identical to the baseline question. Consequently, under sound legal reasoning, the set of applicable legal rules and doctrines governing each variant remains strictly invariant to the introduced descriptor.
- **Systematic Comparative Scope:** Generating ten parallel variants for every baseline question enables comparative evaluation across different demographic and social identities. This breadth allows us to determine whether retrieval shifts are isolated to specific demographic categories or reflect broader systemic sensitivities across different social identities.

Formally, for each original question identifier mbe_i, the dataset comprises ten indexed instances mbe_i_k where $k \in \{1, \ldots, 10\}$, each preserving the underlying ground-truth legal authorities (gold_passage and canonical ground-truth entities) associated with the original unperturbed question.

3.3. Difficulty-Based Stratification

Legal reasoning varies significantly in cognitive complexity, ranging from direct, single-step syllogisms to intricate, multi-layered analyses requiring the synthesis of multiple competing rules, exceptions, and standards. To examine whether demographic sensitivity varies with the legal reasoning complexity of the questions, we classify each question into three difficulty tiers: easy, medium, and hard. The classification is performed by an LLM-based legal-education evaluator using predefined criteria that focus on the amount and complexity of legal reasoning required to reach the correct answer, rather than on question length, vocabulary, or the number of facts.

The classification criteria distinguish the three levels according to the reasoning required to reach the correct answer. Easy questions test a single clear and well-established legal rule or doctrine and generally require only one or two straightforward reasoning steps. Medium questions require applying a legal rule to several facts, multiple meaningful reasoning steps, or distinguishing between related doctrines or rules, potentially in the presence of distracting or conflicting facts. Hard questions require integrating multiple legal rules or doctrines and may involve several sequential reasoning steps, subtle exceptions, competing legal principles, ambiguous facts, or difficult distinctions requiring substantial legal analysis. The classification is therefore based on legal reasoning complexity rather than surface characteristics such as question length.

Each question is submitted to the classification model independently, which is instructed to assign exactly one of the three difficulty labels and return the classification in a predefined JSON format.

Difficulty stratification is necessary because models may exhibit non-uniform sensitivity across cognitive loads. An LLM might consistently retrieve correct governing rules for straightforward questions where salient factual triggers dominate the context, but become susceptible to demographic distraction when tasked with resolving harder, conceptually dense scenarios. Incorporating difficulty-level metadata enables fine-grained cross-stratification, permitting comparisons not only across demographic groups (e.g., `black` versus `white`), but also across difficulty levels within each demographic condition (e.g., easy `black` questions versus hard `black` questions).

3.4. Canonical Legal-Rule Retrieval Task

Rather than evaluating multiple-choice answer selection, our framework probes intermediate legal-rule retrieval. When human legal practitioners analyze a fact pattern, they identify the governing legal concepts, statutory frameworks, and doctrinal tests before applying them to reach a conclusion. Evaluating models at this intermediate stage reveals whether demographic perturbations corrupt the legal basis of reasoning before any final conclusion is reached.

The legal-entity retrieval task is performed in two stages. First, the original unbiased MBE question is provided to the LLM, which is instructed to identify the canonical legal entities that are legally relevant to the fact pattern. These entities constitute the reference set (ground truth) for that question. Second, each corresponding demographic/socially modified version of the same question is provided to the LLM using the same entity-retrieval task, producing the retrieved entity set to be evaluated against the reference set. Because the underlying legally relevant facts are intended to remain unchanged, the entities identified from the unbiased question serve as the reference against which changes in the retrieved legal concepts under demographic perturbation are measured. To avoid superficial token matching, the task explicitly requires the model to abstract away from case-specific proper nouns, party names, and factual descriptions. Instead, models must identify the underlying legal authorities across established canonical entity types, including:

- Core substantive authorities: `LEGAL_DOCTRINE`, `LEGAL_RULE`, `LEGAL_CONCEPT`, `LEGAL_STANDARD`, `LEGAL_CRITERION`, `LEGAL_TEST`, `LEGAL_ELEMENT`;

- Rights, duties, and defenses: `LEGAL_DUTY`, `LEGAL_RIGHT`, `LEGAL_DEFENSE`, `LEGAL_EXCEPTION`, `LEGAL_REMEDY`;
- Statutory and jurisdictional context: `LEGAL_STATUTE`, `LEGAL_CASE`, `LEGAL_CITATION`, `LEGAL_REGULATION`, `LEGAL_JURISDICTION`, `LEGAL_COURT`; and
- Procedural and transactional mechanisms: `LEGAL_PROCEDURE`, `LEGAL_EVENT`, `LEGAL_OFFER`, `LEGAL_CONTRACT_TERM`, `LEGAL_PARTY`, `LEGAL_ACTOR`.

Models are instructed to return the 5 to 10 most relevant canonical legal entities governing the fact pattern. To enforce structural consistency and ensure automated evaluation without human parsing ambiguity, the model's output is constrained to a predefined JSON schema specifying an array of entity objects, each containing the canonical `name` of the legal concept and its corresponding categorical `type`.

3.5. Accounting for Generation Variability via Repeated Sampling

Because LLM generation is subject to stochastic sampling variability, a single inference run per prompt cannot distinguish between random generation noise and systematic demographic sensitivity. If a model retrieves an erroneous legal rule for a `"black"` variant in a single run, that divergence could be an artifact of decoding variance rather than a repeatable sensitivity to the demographic descriptor.

To resolve this ambiguity, our framework incorporates a repeated-generation protocol. For every demographic variant `mbe_i_k`, we execute $N = 5$ independent retrieval runs under identical prompt conditions.

This protocol serves two methodological purposes:

1. **Separating Stochasticity from Systematic Shift:** By observing the distribution of retrieved legal entities over repeated generations, we can verify whether demographic divergences persist across runs or represent isolated sampling fluctuations.
2. **Measuring Retrieval Stability:** Repeated sampling enables measuring intra-condition variance. A model that consistently retrieves the same legal rules across all five runs of a baseline or demographic variant demonstrates higher doctrinal stability than one whose retrieved authorities fluctuate across runs.

3.6. Semantic Ground-Truth Alignment and Evaluation Framework

Evaluating retrieved legal rules presents a distinct semantic challenge: legal rules, doctrines, and concepts can be expressed through diverse phrasing, synonymous terminology, or varying levels of syntactic granularity (e.g., *impeachment of witness* versus *witness impeachment for prior inconsistent statement*). Consequently, exact lexical matching severely under-measures retrieval quality.

To evaluate retrieval fidelity accurately, we implement a semantic alignment framework that pairs predicted entities with canonical ground-truth entities through dense vector representations, global bipartite matching, and similarity thresholding.

3.6.1. Ground-Truth Extraction and Embedding

For each original unbiased question, the LLM is first prompted to identify the legally relevant canonical entities using the same structured entity schema employed for the evaluation queries. The resulting entity set is treated as the reference ground truth $\mathcal{G}$. For each demographic/social variant, the LLM independently performs the same entity-retrieval task, producing a predicted entity set $\mathcal{P}$. Thus, the evaluation compares the legal entities identified from the biased variant against those identified from its corresponding unbiased question. For any demographic variant derived from For any demographic variant derived from mbe_i, the model outputs a set of predicted legal entities $\mathcal{P} = \{p_1, p_2, \ldots, p_{|\mathcal{P}|}\}$ for each independent run.

Every canonical entity name in $\mathcal{G}$ and $\mathcal{P}$ is mapped into a high-dimensional semantic embedding space using a dedicated legal embedding model:

$$\mathbf{v}_g = \text{Embed}(g), \quad \mathbf{v}_p = \text{Embed}(p)$$

All computed embeddings are cached in a structured database to ensure reproducibility, deterministic metric computation, and computational efficiency across iterations.

3.6.2. Optimal Bipartite Matching

To determine the correspondence between the predicted entities $\mathcal{P}$ and the ground-truth entities $\mathcal{G}$, we construct a pairwise cosine similarity matrix $\mathbf{S} \in \mathbb{R}^{|\mathcal{P}|\times|\mathcal{G}|}$, where:

$$\mathbf{S}_{j,l} = \frac{\mathbf{v}_{p_j} \cdot \mathbf{v}_{g_l}}{|\mathbf{v}_{p_j}||\mathbf{v}_{g_l}|} \tag{1}$$

We treat the alignment between $\mathcal{P}$ and $\mathcal{G}$ as a maximum-weight bipartite matching problem, equivalent to minimizing an assignment cost matrix $\mathbf{C} = \mathbf{1} - \mathbf{S}$. We solve for the global optimal one-to-one matching using the Hungarian assignment algorithm (the Kuhn-Munkres algorithm):

$$\pi^* = \arg\min_{\pi} \sum_{j} \mathbf{C}_{j,\pi(j)} \tag{2}$$

This assignment guarantees that each predicted entity is matched to at most one ground-truth entity in a manner that globally maximizes semantic similarity, precluding greedy or double-counted alignments.

3.6.3. Thresholding and Retrieval Metrics

For each matched pair $(p_j, g_{\pi^*(j)})$, we evaluate whether the semantic similarity exceeds a predefined threshold τ:

$$\text{sim}(p_j, g_{\pi^*(j)}) \geq \tau \tag{3}$$

Rather than fixing the threshold to a single value, we evaluate retrieval performance across five similarity thresholds: $\tau \in \{0.70, 0.75, 0.80, 0.85, 0.90\}$.

For each threshold, a matched pair satisfying the corresponding similarity criterion is considered a valid retrieval of the ground-truth legal entity. This allows the evaluation to examine retrieval performance under different levels of semantic similarity strictness.

Based on each threshold, we categorize the retrieval outcomes as follows:

- **True Positives (TP):** The number of matched entity pairs (p_j, g_l) whose semantic similarity is greater than or equal to the specified threshold.
- **False Positives (FP):** Predicted entities that either remain unassigned or whose optimal match fails to reach the specified threshold, such that $FP = |\mathcal{P}| - TP$.
- **False Negatives (FN):** Ground-truth entities that are not matched to any predicted entity with similarity greater than or equal to the specified threshold, such that $FN = |\mathcal{G}| - TP$.

For each threshold, we then compute Precision, Recall, and F1:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{4}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{5}$$

$$\text{F1} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{6}$$

Thus, retrieval performance is evaluated separately at each of the five similarity thresholds, allowing the effect of the semantic matching criterion on retrieval quality to be examined.

4. Implementation

This section details the concrete computational pipeline and experimental architecture used to implement the evaluation framework described in Section 3. The pipeline operationalizes the legal-rule retrieval task, manages asynchronous model inference across repeated sampling runs, and executes semantic ground-truth alignment with bipartite matching and persistent embedding caching.

4.1. Dataset Preparation and Experimental Stratification

The experimental dataset is instantiated from a structured tabular repository containing the perturbed Multistate Bar Examination (MBE) instances `biased_final2.csv`. Each record represents an experimental item indexed by an identifier `mbe_`i`_`k, where i denotes the original baseline MBE fact pattern and $k \in \{1, \ldots, 10\}$ corresponds to one of the ten demographic or social conditions:

1. `black`
2. `white`
3. `arab`
4. `asian`
5. `hispanic`
6. `wealthy`
7. `low-income`
8. `homeless`
9. `immigrant`
10. `female`

The text submitted to the language models is extracted from the `biased_question` column, which embeds the legally irrelevant demographic perturbation into the factual narrative while strictly preserving the underlying operative facts and legal questions.

To operationalize the complexity dimension, difficulty annotations are loaded from an external classification file `create_category2.csv`. This file maps each baseline question identifier mbe_i directly to one of the three established categorical tiers (`easy`, `medium`, or `hard`), enabling joint cross-stratification with demographic variants during post-processing and metric aggregation.

4.2. Canonical Legal-Rule Extraction Pipeline

The legal-rule retrieval queries are submitted to large language models through structured inference pipelines. The experimental framework evaluates five contemporary models:

1. `gemini-3.1-flash-lite`
2. `gpt-5-mini`
3. `claude-haiku-4.5`
4. `llama-4-scout`
5. `llama-4-maverick`

4.2.1. Prompt Configuration

Model prompting is standardized across all conditions. The prompt directs the language model to act as an extraction system generating retrieval queries for a legal knowledge graph constructed from explanatory legal texts rather than raw fact patterns. The instructions explicitly instruct the model to abstract away from specific party names and superficial factual tokens, requiring instead the extraction of 5 to 10 governing canonical legal nodes.

The prompt defines 24 canonical node types:

- `LEGAL_DOCTRINE`
- `LEGAL_RULE`
- `LEGAL_CONCEPT`
- `LEGAL_CRITERION`
- `LEGAL_STANDARD`
- `LEGAL_DUTY`
- `LEGAL_RIGHT`
- `LEGAL_EXCEPTION`
- `LEGAL_DEFENSE`
- `LEGAL_PROCEDURE`
- `LEGAL_EVENT`

- LEGAL_PARTY
- LEGAL_ACTOR
- LEGAL_CITATION
- LEGAL_CASE
- LEGAL_STATUTE
- LEGAL_TEST
- LEGAL_ELEMENT
- LEGAL_REMEDY
- LEGAL_OFFER
- LEGAL_CONTRACT_TERM
- LEGAL_JURISDICTION
- LEGAL_COURT
- LEGAL_REGULATION

4.2.2. Structured Output Enforcement

To guarantee parseability and eliminate ambiguities in entity identification, outputs are constrained via strict JSON schema validation. The API request configuration enforces a schema requiring an outer object containing an `entities` array. Each item within the array is constrained to an object with exactly two required properties:

- `name`: a string specifying the canonical legal concept; and
- `type`: a closed string enumeration restricted strictly to the 24 predefined canonical node types.

The schema disables additional object properties and enables strict schema compliance (`strict: True`), ensuring that model outputs conform exactly to the specified data structure without extraneous conversational wrapper tokens.

4.3. Repeated Generation and Asynchronous Batch Execution

To account for stochastic generation variability, each demographic variant `mbe_i_k` is evaluated across $N = 5$ independent retrieval runs.

4.3.1. Asynchronous Batch Management

For API-based inference, requests are organized and dispatched asynchronously using the Batch API endpoint. Requests are batched into chunks of up to 5,000 items per submission. Each individual query within a batch is tagged with a unique, sanitized tracking identifier adhering to the format:

$$\texttt{custom_id} = \texttt{sanitized_idx__run_run_number} \tag{7}$$

where $\texttt{run_number} \in \{0, 1, 2, 3, 4\}$

4.3.2. Fault Tolerance and State Persistence

The generation harness maintains state persistence to ensure resilience against process interruptions, network drops, or rate limits:

- **Batch State Tracking:** An on-disk state file logs active batch identifiers, submitted request payloads, and retry counters. The pipeline polls the API at 60-second intervals to monitor execution status (`validating`, `in_progress`, `finalizing`, `completed`, `failed`, `expired`, `cancelled`).
- **Incremental Execution:** Before submitting requests, the system reads existing generations from an append-only JSONL file. A lookup table checks how many completed runs exist for each identifier; only instances with fewer than five successful generations are queued for execution.
- **Error Handling and Retry Logic:** If individual requests within a completed batch fail (e.g., non-200 HTTP responses, parsing exceptions, or malformed JSON payloads), or if an entire batch fails or expires, the failed requests are isolated and automatically resubmitted across up to 20 retry cycles (`MAX_RETRY_BATCHES = 20`).
- **Atomic File Updates:** All dataset writing routines write to temporary files (`.tmp`) and perform atomic filesystem replacements to prevent data corruption during write operations.

4.4. Dense Embedding Generation and SQLite Caching

Evaluating retrieved entities against reference sets requires projecting all legal entity names into a dense vector space.

4.4.1. Embedding Model

Entity embeddings are generated using `gemini-embedding-2` model via the Google GenAI SDK. The model is configured to produce vectors with a fixed dimensionality of:

$$\mathbf{v} \in \mathbb{R}^{1536}$$

via the `output_dimensionality` parameter.

4.4.2. SQLite Cache Architecture

To prevent redundant API calls across multiple models, runs, and evaluation sweeps, all embeddings are stored in a centralized SQLite database (`embedding_cache.db`). The database schema maintains a dedicated table where embedding vectors are serialized directly as raw 32-bit floating-point byte buffers.

4.4.3. Concurrent Generation and Rate Limiting

Prior to evaluation, the pipeline aggregates all unique entity strings across both ground-truth definitions and model predictions. Entities absent from the database are isolated and queued for batched embedding:

- **Batching and Multithreading:** Missing entity names are partitioned into batches of 100 items and submitted concurrently using a `ThreadPoolExecutor` configured with 5 worker threads.
- **Global Rate Limiting:** A process-wide mutual exclusion lock enforces a strict minimum delay of 2.0 seconds between outgoing API calls across all threads.
- **Exponential Backoff:** If an embedding request encounters an error, the batch retries up to 5 times using exponential backoff with additive random uniform jitter:

$$\Delta t = 2^{\text{attempt}} + \text{Uniform}(0, 1.0) \tag{8}$$

4.5. Bipartite Alignment, Metric Computation, and Export

The evaluation harness processes model predictions sequentially across all five model configurations using their corresponding ground-truth reference files and prediction files.

4.5.1. Semantic Matching Implementation

For each baseline question `mbe_`i and variant `mbe_i_k`, ground-truth entities $\mathcal{G} = g_1, \ldots, g_{|\mathcal{G}|}$ and predicted entities $\mathcal{P} = p_1, \ldots, p_{|\mathcal{P}|}$ for run $r \in 0, \ldots, 4$ are extracted and stripped of whitespace. If either $\mathcal{G}$ or $\mathcal{P}$ is empty:

$$TP = 0, \quad FP = |\mathcal{P}|, \quad FN = |\mathcal{G}| \tag{9}$$

yielding Precision, Recall, and F1 scores of 0. When both sets contain valid entities, their dense vectors are retrieved from the SQLite database:

1. **Cosine Similarity Matrix:** The pairwise similarity matrix $\mathbf{S} \in \mathbb{R}^{|\mathcal{P}|\times|\mathcal{G}|}$ is computed via `sklearn.metrics`:

$$\mathbf{S}_{j,l} = \frac{\mathbf{v}_{p_j} \cdot \mathbf{v}_{g_l}}{|\mathbf{v}_{p_j}||\mathbf{v}_{g_l}|} \tag{10}$$

2. **Cost Matrix Formulation:** A cost matrix is constructed as:

$$\mathbf{C} = 1 - \mathbf{S} \tag{11}$$

3. **Linear Sum Assignment:** The optimal global bipartite match is computed using the Kuhn-Munkres algorithm implemented in `scipy.optimize`:

$$(\mathbf{p}, \mathbf{g}) = \texttt{linear_sum_assignment}(\mathbf{C}) \tag{12}$$

4. **Threshold Filtering:** The operational evaluation applies a similarity threshold of 0.70, 0.75, 0.80, 0.85 and 0.90, for example if:

$$\tau = 0.70 \tag{13}$$

For every assigned index pair (p_m, g_m), the assignment is counted as a true positive if:

$$\mathbf{S}_{p_m,g_m} \geq 0.70 \tag{14}$$

4.5.2. Metric Derivation

From the assigned matches, retrieval statistics are calculated as:

$$TP = \sum_m \mathbb{I}\left(\mathbf{S}_{p_m,g_m} \geq 0.70\right) \tag{15}$$

$$FP = |\mathcal{P}| - TP \tag{16}$$

$$FN = |\mathcal{G}| - TP \tag{17}$$

Precision, Recall, and F1 are derived directly:

$$\text{Precision} = \begin{cases} \frac{TP}{TP+FP} & \text{if } TP + FP > 0 \\ \text{otherwise} & 0 \end{cases} \tag{18}$$

$$\text{Recall} = \begin{cases} \frac{TP}{TP+FN} & \text{if } TP + FN > 0 \\ \text{otherwise} & 0 \end{cases} \tag{19}$$

$$\text{F1} = \begin{cases} \frac{2\cdot\text{Precision}\cdot\text{Recall}}{\text{Precision}+\text{Recall}} & \text{if Precision} + \text{Recall} > 0 \\ \text{otherwise} & 0 \end{cases} \tag{20}$$

4.5.3. Output Formatting and Stateful Resumption

Evaluation results are exported to tab-separated value (TSV) files. The exported schema records 18 attributes per run:

- **Identifier and stratification metadata:** `mbe_id`, `idx`, `biased_category`, `difficulty`, `model`, `threshold`, `temperature`, and `run`;
- **Entity counts:** `gt_count` and `pred_count`;
- **Confusion metrics:** `tp`, `fp`, and `fn`;
- **Retrieval performance:** `precision`, `recall`, and `f1` (formatted to four decimal places); and
- **Raw entity strings:** `gt_entities` and `pred_entities`

The evaluation script supports stateful resumption. Existing output TSVs are read upon startup, populating a set of completed tuples (`mbe_id`, `idx`, `run`). If an execution is interrupted, existing records are bypassed, ensuring idempotent and efficient re-computation.

5. Results and Discussion

This section evaluates whether introducing legally irrelevant demographic and social descriptors into identical legal fact patterns alters retrieved legal rules and degrades retrieval consistency. The experimental design comprises 1,045 Multistate Bar Examination (MBE) fact patterns, each perturbed across 10 demographic and socioeconomic categories: arab, asian, black, female, hispanic, homeless, immigrant, low-income, wealthy, and white. These conditions

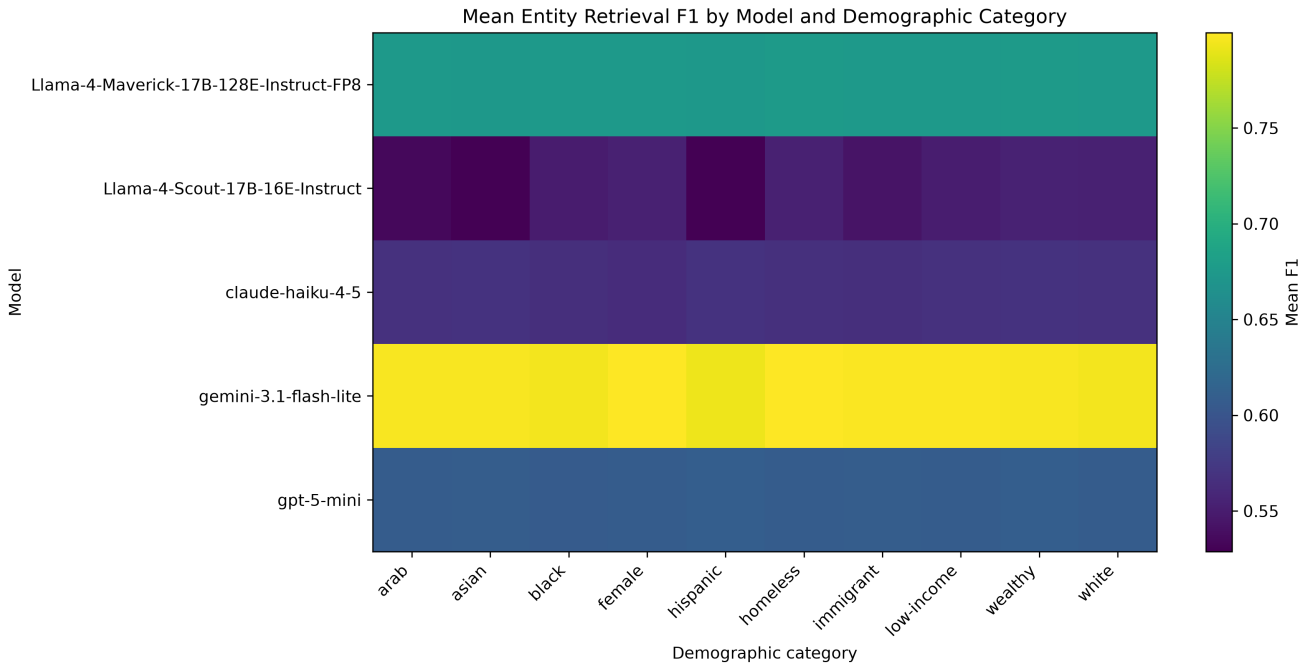


Figure 2: Mean Entity Retrieval F1 by Model and Demographic Category.

were evaluated across five large language models: gemini-3.1-flash-lite, Llama-4-Maverick, gpt-5-mini, claude-haiku-4-5, and Llama-4-Scout. Retrieved legal rules were evaluated against reference target entities under five Hungarian bipartite matching semantic similarity thresholds $\tau \in \{0.70, 0.75, 0.80, 0.85, 0.90\}$ across five repeated stochastic runs, yielding 1,306,000 raw evaluation instances and 250,000 run-averaged condition-level observations.

5.1. Overall Retrieval Performance and Precision-Recall Regimes

Across the experimental corpus, the evaluated architectures exhibit marked variations in baseline retrieval accuracy, precision-recall trade-offs, and output entity volumes. Figure 2 presents the mean Entity Retrieval F1 scores across the five evaluated models and ten demographic categories.

The heatmap shows a clear and consistent difference in retrieval performance between models, with gemini-3.1-flash-lite achieving the highest mean F1 scores across all demographic categories, followed by Llama-4-Maverick. In contrast, Llama-4-Scout exhibits the lowest mean F1 scores, while gpt-5-mini and claude-haiku-4.5 occupy intermediate ranges. Notably, variation across demographic categories within each model is comparatively small relative to the differences between models, suggesting that model-level differences are more pronounced than demographic differences in this visualization. A small amount of within-model variation is nevertheless visible, particularly for Llama-4-Scout, where the Asian and Hispanic categories show slightly lower F1 values than several other categories. Overall, the heatmap indicates that retrieval performance is primarily differentiated by model, while demographic variation appears comparatively modest. However, because the figure presents aggregated means without measures of variance or statistical significance, the observed differences between demographic categories should not, by themselves, be interpreted as evidence of statistically significant demographic disparities.

Table 1 reports the macro- and micro-level retrieval performance metrics aggregated across all demographic variants, matching thresholds, and repeated runs. gemini-3.1-flash-lite achieved the highest overall retrieval accuracy, exhibiting a balanced operational profile across precision (0.804642) and recall (0.793806) alongside the lowest overall dispersion (standard deviation of F1 = 0.147133). In contrast, Llama-4-Maverick operated in a recall-dominant regime, attaining a recall of 0.754037 against a lower precision of 0.639403. Both gpt-5-mini and claude-haiku-4-5 exhibited highly conservative, precision-dominant behaviors characterized by high entity omission: gpt-5-mini attained the highest precision in the benchmark (0.867208) while retrieving less than half of the target legal entities (recall = 0.480139). Similarly, claude-haiku-4-5 recorded a precision of 0.789746 against a recall of 0.451172. Llama-4-Scout recorded the lowest overall retrieval accuracy (macro F1 = 0.544958) and the highest score dispersion (standard deviation = 0.240381).

The observed divergence between macro- and micro-averaged F1 scores illustrates volume-dependent retrieval behavior on questions with dense ground-truth entity targets. For gpt-5-mini, macro F1 (0.608248) exceeded micro F1 (0.594792), whereas for Llama-4-Scout, micro F1 (0.553677) exceeded macro F1 (0.544958).

Table 1
Overall retrieval performance by model across 1,306,000 evaluations.

Model	Macro F1	Micro F1	Macro Precision	Macro Recall	SD (F1)
gemini-3.1-flash-lite	0.796955	0.797051	0.804642	0.793806	0.147133
Llama-4-Maverick-17B-128E-Instruct-FP8	0.674918	0.679216	0.639403	0.754037	0.179861
gpt-5-mini	0.608248	0.594792	0.867208	0.480139	0.168959
claude-haiku-4-5	0.566678	0.568690	0.789746	0.451172	0.189547
Llama-4-Scout-17B-16E-Instruct	0.544958	0.553677	0.553682	0.569501	0.240381

Table 2
Highest and lowest scoring demographic categories by model.

Model	Highest Category (F1)	Lowest Category (F1)	Demographic F1 Gap
gemini-3.1-flash-lite	female (0.7998)	hispanic (0.7934)	0.006468
Llama-4-Maverick-17B-128E-Instruct-FP8	wealthy (0.6769)	hispanic (0.6732)	0.003772
gpt-5-mini	hispanic (0.6098)	black (0.6071)	0.002767
claude-haiku-4-5	wealthy (0.5686)	female (0.5636)	0.005005
Llama-4-Scout-17B-16E-Instruct	wealthy (0.5545)	hispanic (0.5287)	0.025787

5.2. Macro Demographic Parity and Directional Heterogeneity

At the macro population level, demographic category substitution did not produce systematic performance suppression across protected groups. Aggregated across all models, thresholds, and runs (1,306,000 evaluations), retrieval metrics remained tightly bounded across all 10 demographic categories. Aggregate precision ranged from 0.7270 (hispanic) to 0.7342 (wealthy), a spread of 0.0072. Aggregate recall ranged from 0.6064 (hispanic) to 0.6125 (wealthy), a spread of 0.0061. Aggregate F1 spanned strictly between 0.6346 (hispanic) and 0.6413 (wealthy), representing an aggregate cross-category spread of 0.0067 F1 points. Error counts exhibited identical parity across conditions: mean true positives ranged from 6.337 entities per prompt (hispanic) to 6.387 entities (wealthy), mean false positives ranged from 2.338 (hispanic) to 2.367 (low-income), and mean false negatives spanned from 5.016 (wealthy) to 5.066 (hispanic), against an exact ground-truth benchmark mean of 11.4029 entities across all demographic conditions.

Furthermore, descriptive aggregate performance rankings did not mirror real-world socioeconomic stratification. Fact pattern variants specifying socioeconomically vulnerable groups yielded aggregate F1 scores exceeding those of baseline majority categories: homeless fact patterns achieved an aggregate F1 of 0.6406 and low-income variants achieved 0.6398, both exceeding white (0.6393), black (0.6385), asian (0.6354), and hispanic (0.6346). In paired testing across 1,045 questions, the difference between homeless and wealthy was statistically indistinguishable (mean difference = 0.00067, $p = 0.3614$).

Directional sensitivity across categories varied substantially by model architecture, as detailed in Table 2. A demographic category disfavored by one architecture was frequently favored by another.

While hispanic variants recorded the lowest F1 in Llama-4-Maverick (0.6732), Llama-4-Scout (0.5287), and gemini-3.1-flash-lite (0.7934), hispanic variants achieved the highest F1 in gpt-5-mini (0.6098). Similarly, female variants recorded the lowest performance in claude-haiku-4-5 (0.5636) but achieved the highest performance in gemini-3.1-flash-lite (0.7998). This directional heterogeneity confirms that demographic variation in retrieval performance does not reflect a consistent, cross-architectural penalty against specific groups.

5.3. Mixed-Effects Regression and Architectural Divergence

To formally assess the contributions of demographic categories, model architectures, semantic thresholds, and question difficulty, a linear mixed-effects model was fitted on 250,000 run-averaged condition rows:

$$\begin{aligned} \mathrm{F1}_{\mathrm{mean}} \sim\ & \mathrm{biased_category} + \mathrm{model} + \mathrm{threshold} + \mathrm{difficulty} \\ & + (\mathrm{biased_category} \times \mathrm{model}) \\ & + (\mathrm{biased_category} \times \mathrm{threshold}) \end{aligned}$$

with random intercepts specified for legal questions (mbe_id) and variance components for observation index (idx). In this parameterization, arab served as the baseline reference category for demographic descriptors, Llama-4-Maverick served as the reference model, threshold 0.70 as the reference matching threshold, and easy as the reference difficulty level.

Table 3 reports the fixed-effect estimates for model architectures and question difficulty tiers.

Model fixed effects were highly significant across all architectures relative to Llama-4-Maverick ($p <$ 1.0e-158). Conversely, demographic category main effects relative to the reference category (arab) were uniformly null across the corpus, with all standard errors equal to 0.003352 and p-values exceeding 0.50:

Table 3
Linear mixed-effects regression coefficients for model and difficulty main effects.

Factor / Term	Coefficient (β)	SE	z-value	p-value	95% Confidence Interval
Model (Ref: Llama-4-Maverick)					
gemini-3.1-flash-lite	+0.122394	0.002495	+49.056	< 1.0e-300	[0.117504, 0.127283]
gpt-5-mini	-0.066970	0.002495	-26.842	1.05e-158	[-0.071859, -0.062080]
claude-haiku-4-5	-0.108213	0.002495	-43.373	< 1.0e-300	[-0.113103, -0.103324]
Llama-4-Scout-17B-16E-Instruct	-0.140636	0.002496	-56.354	< 1.0e-300	[-0.145527, -0.135744]
Difficulty (Ref: Easy)					
medium	+0.013521	0.015312	+0.883	0.3772	[-0.016490, 0.043532]
hard	+0.010752	0.017181	+0.626	0.5314	[-0.022922, 0.044425]

- **asian:** $\beta = -0.001391$, $z = -0.415$, $p = 0.678267$, CI : $[-0.007961, 0.005179]$
- **black:** $\beta = 0.001240$, $z = 0.370$, $p = 0.711507$, CI : $[-0.005330, 0.007810]$
- **female:** $\beta = 0.001926$, $z = 0.575$, $p = 0.565663$, CI : $[-0.004644, 0.008496]$
- **hispanic:** $\beta = -0.002227$, $z = -0.664$, $p = 0.506593$, CI : $[-0.008797, 0.004343]$
- **homeless:** $\beta = 0.001876$, $z = 0.560$, $p = 0.575776$, CI : $[-0.004694, 0.008446]$
- **immigrant:** $\beta = -0.000363$, $z = -0.108$, $p = 0.913827$, CI : $[-0.006933, 0.006207]$
- **low-income:** $\beta = 0.000977$, $z = 0.292$, $p = 0.770623$, CI : $[-0.005593, 0.007547]$
- **wealthy:** $\beta = 0.001959$, $z = 0.584$, $p = 0.558956$, CI : $[-0.004611, 0.008529]$
- **white:** $\beta = -0.000127$, $z = -0.038$, $p = 0.969843$, CI : $[-0.006697, 0.006443]$

A substantial architectural divergence appeared in the demographic interaction terms. For gemini-3.1-flash-lite, claude-haiku-4-5, and gpt-5-mini, all demographic-by-model interaction terms were non-significant (gemini: all $p \in [0.4010, 0.9235]$; claude: all $p \in [0.2415, 0.9392]$; gpt-5-mini: all $p \in [0.3184, 0.9127]$). In contrast, Llama-4-Scout demonstrated statistically significant positive interactions across 7 of 9 demographic categories relative to the reference category (arab): white ($\beta = 0.020082, z = 5.690, p =$ 1.27e-08), homeless ($\beta = 0.018262, z = 5.175, p =$ 2.28e-07), wealthy ($\beta = 0.018002, z = 5.101, p =$ 3.38e-07), female ($\beta = 0.017985, z = 5.096, p =$ 3.47e-07), low-income ($\beta = 0.015337, z = 4.346, p =$ 1.39e-05), black ($\beta = 0.014814, z = 4.197, p =$ 2.70e-05), and immigrant ($\beta = 0.009299, z = 2.635, p =$ 0.00842). The interactions in Scout for asian ($\beta = -0.004583, z = -1.299, p =$ 0.194124) and hispanic ($\beta = -0.003964, z = -1.123, p =$ 0.261395) remained non-significant.

Non-parametric Friedman repeated-measures omnibus tests across the 10 demographic categories confirmed this architectural split. As reported in Table 4, four of five models exhibited omnibus demographic rank invariance, whereas Llama-4-Scout exhibited statistically significant rank instability.

Scaling the mixture-of-experts architecture from 16 experts (Llama-4-Scout) to 128 experts (Llama-4-Maverick) while maintaining active parameter scale at 17B eliminated demographic omnibus rank instability: the Friedman statistic dropped from 69.58 ($p =$ 1.84e-11) to 10.18 ($p =$ 0.3358), mean within-question demographic F1 range decreased from 0.2019 to 0.1278, and maximum demographic range contracted from 0.8795 to 0.5156.

5.4. Intra-Question Instability and Difficulty Dynamics

In the mixed-effects regression, legal question difficulty tiers did not yield statistically significant main effects on condition F1 relative to easy questions (medium: $\beta = 0.013521, p = 0.3772$; hard: $\beta = 0.010752, p = 0.5314$). However, difficulty tiers were associated with a descriptive operational rebalancing between precision and recall across models. As difficulty scaled from easy to hard, precision increased monotonically ($0.6775 \rightarrow 0.7304 \rightarrow 0.7523$) while recall dropped monotonically ($0.6291 \rightarrow 0.6110 \rightarrow 0.5963$). In more doctrinally intricate cases, retrieval behavior aligned with a more conservative profile featuring higher precision and lower recall.

Figure 3 illustrates Entity Retrieval Performance, measured by Mean F1, across easy, medium, and hard questions for the five evaluated models. `gemini-3.1-flash-lite` consistently achieves the highest retrieval performance across all difficulty levels, with Mean F1 remaining close to 0.80 (0.796 for easy, 0.795 for medium, and 0.806 for hard). `Llama-4-Maverick-17B-128E-Instruct-FP8` shows an increasing trend as difficulty rises, from approximately 0.622 for easy questions to 0.676 for medium and 0.687 for hard questions. In contrast, `gpt-5-mini` decreases from approximately 0.638 on easy questions to 0.606 on medium questions, after which performance remains largely stable. `claude-haiku-4-5` also remains relatively stable between easy and medium questions before decreasing to approximately 0.538 on hard questions, while `Llama-4-Scout-17B-16E-Instruct` increases from approximately 0.506 to 0.548 before a slight decline to 0.542. These results indicate that question difficulty does not produce a uniform decrease in entity retrieval performance across models. Instead, the direction and magnitude

Table 4
Friedman omnibus tests of demographic rank invariance by model.

Model	Evaluated Questions (n)	Friedman Statistic (χ^2)	p-value	Significance Status
gemini-3.1-flash-lite	1,045	13.80	0.1297	Not significant ($p > 0.05$)
Llama-4-Maverick	1,045	10.18	0.3358	Not significant ($p > 0.05$)
gpt-5-mini	1,045	15.01	0.0906	Not significant ($p > 0.05$)
claude-haiku-4-5	1,045	15.93	0.0684	Not significant ($p > 0.05$)
Llama-4-Scout	1,044	69.58	1.84e-11	Statistically significant

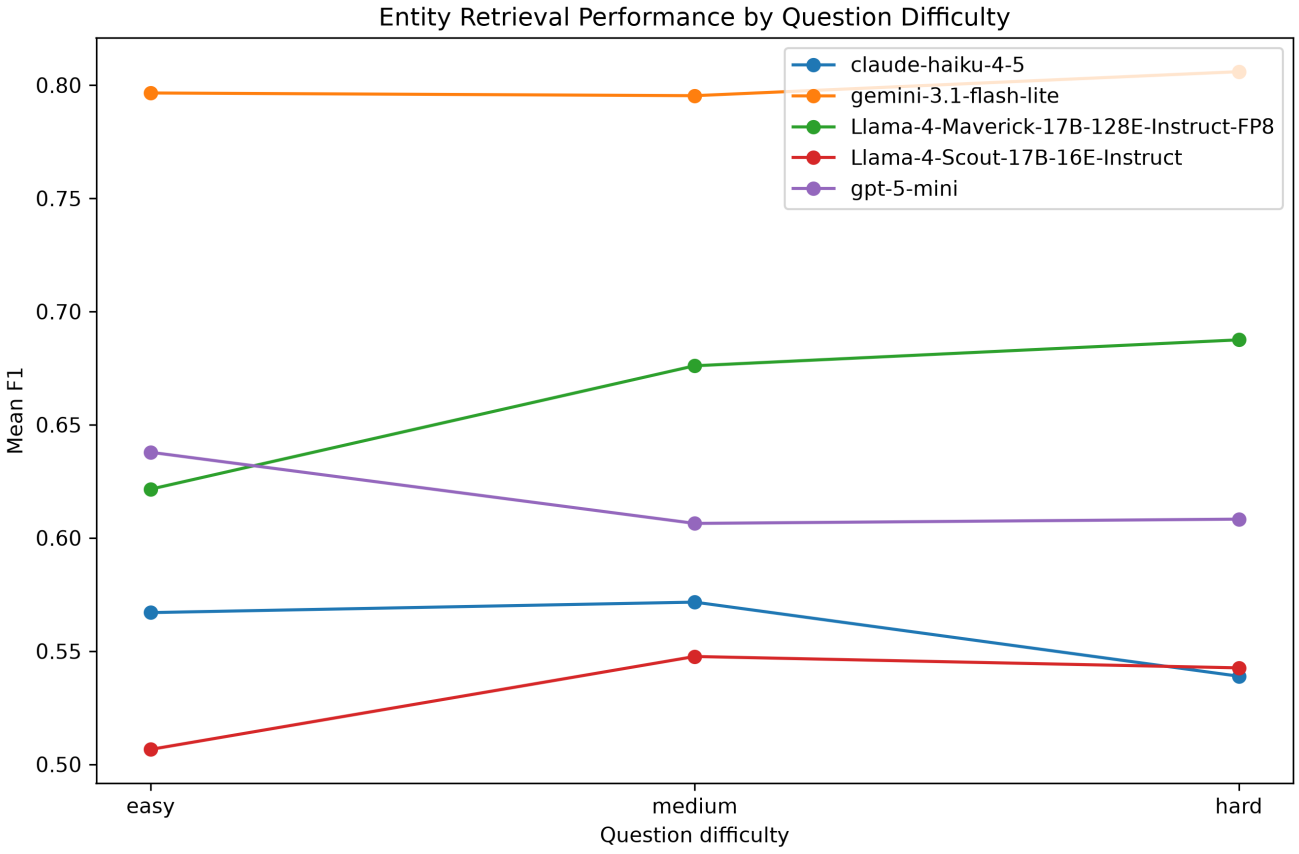


Figure 3: Entity Retrieval Performance by Question Difficulty.

of changes vary by model, with some models improving as difficulty increases and others showing stable or declining performance. As the figure does not include measures of variance or statistical significance, these visual differences should be interpreted as descriptive rather than evidence of statistically significant effects.

At the item level, counterfactual perturbations of identical fact patterns produced extensive retrieval variance. Despite macro-level parity, 99.21% (25,889 of 26,095) of complete question-model-threshold profiles displayed non-zero F1 dispersion across demographic variants. Within-question demographic F1 ranges (defined as max(F1) − min(F1) across the 10 variants for a given question condition) exhibited right-skewed distributions across all models, as reported in Table 5.

Among questions displaying high variability, variant F1 ranges spanned between 0.6398 and 1.0. For example, in gemini-3.1-flash-lite on question mbe_198 at threshold 0.80, variant F1 spanned from 0.0 to 1.0 across counterfactual variants. However, this divergence reflects the compounding interplay of prompt alteration, high item difficulty, and stochastic sampling dispersion rather than an isolated deterministic demographic effect. On question mbe_871 in Llama-4-Scout at threshold 0.70, F1 varied from 0.0 to 0.8795 across counterfactual demographic variants.

5.5. Semantic Threshold Sensitivity and Rank Reversals

Evaluating retrievals across Hungarian bipartite matching thresholds from $\tau = 0.70$ to $\tau = 0.90$ revealed monotonic performance degradation across all architectures as visible in Figure 4. In the linear mixed-effects model shown in Table 6, all threshold increments produced statistically

Table 5
Within-question demographic F1 range distributions across models.

Model	Mean Range	Median Range	SD (Range)	Maximum Range
Llama-4-Scout-17B-16E-Instruct	0.2019	0.1667	0.1472	0.8795
gemini-3.1-flash-lite	0.1278	0.1000	0.1134	1.0000
Llama-4-Maverick-17B-128E-Instruct-FP8	0.1278	0.1000	0.1062	0.5156
claude-haiku-4-5	0.1150	0.0909	0.1045	0.5402
gpt-5-mini	0.0894	0.0714	0.0833	0.4682

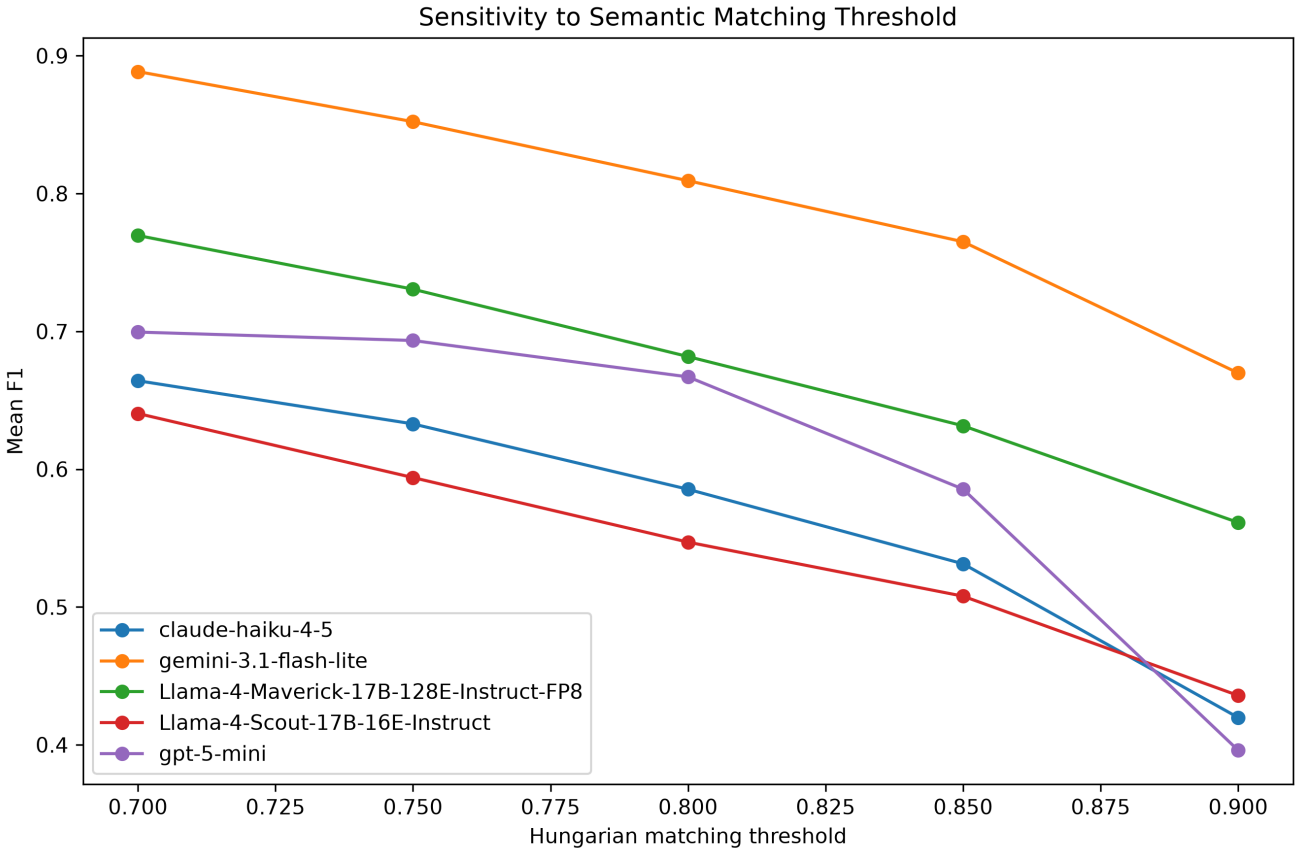


Figure 4: Sensitivity to Semantic Matching Threshold.

significant negative fixed effects relative to $\tau = 0.70$ ($p <$ 1.0e-35).

F1 degradation from threshold 0.70 to 0.90 totaled -0.204660 in Llama-4-Scout (0.640388 → 0.435728), -0.208305 in Llama-4-Maverick (0.769623 → 0.561318), -0.218788 in gemini-3.1-flash-lite (0.888494 → 0.669706), -0.244456 in claude-haiku-4-5 (0.664184 → 0.419728), and -0.303536 in gpt-5-mini (0.699479 → 0.395943).

From threshold 0.70 to 0.85, the model hierarchy was strictly preserved. At threshold 0.90, gpt-5-mini exhibited an abrupt decline: between thresholds 0.85 and 0.90, its F1 fell by -0.1895 while precision fell from 0.8377 to 0.5658. This drop induced a rank reversal at threshold 0.90, where Llama-4-Scout (0.435728) and claude-haiku-4-5 (0.419728) both surpassed gpt-5-mini (0.395943).

All 36 demographic-by-threshold interaction terms in the mixed-effects regression were non-significant ($p \in [0.5269, 0.9685]$). Increasing semantic matching stringency degraded retrieval performance across all demographic groups uniformly, without magnifying or compressing demographic differences.

5.6. Stochastic Run-to-Run Dispersion

To evaluate baseline generation stability independent of demographic substitution, five repeated runs were conducted under identical conditions (identical question, demographic variant, model, and threshold).

Figure 5 presents the distribution of Mean F1 across repeated runs for the five evaluated models, highlighting substantial variability across the evaluated instances. `gemini-3.1-flash-lite` exhibits the highest central tendency, with an interquartile range (IQR) of approximately 0.73–0.89, while `Llama-4-Maverick-17B-128E-Instruct-FP8` has a median of approximately 0.70 and an IQR of 0.59–0.79.

Table 6
Linear mixed-effects regression coefficients for semantic matching thresholds.

Threshold	Coefficient (β)	SE	z-value	p-value	95% Confidence Interval
0.75	-0.031421	0.002495	-12.592	2.33e-36	[-0.036312, -0.026531]
0.80	-0.073988	0.002495	-29.652	3.20e-193	[-0.078878, -0.069098]
0.85	-0.128285	0.002495	-51.413	< 1.0e-300	[-0.133175, -0.123394]
0.90	-0.235105	0.002495	-94.223	< 1.0e-300	[-0.239995, -0.230214]

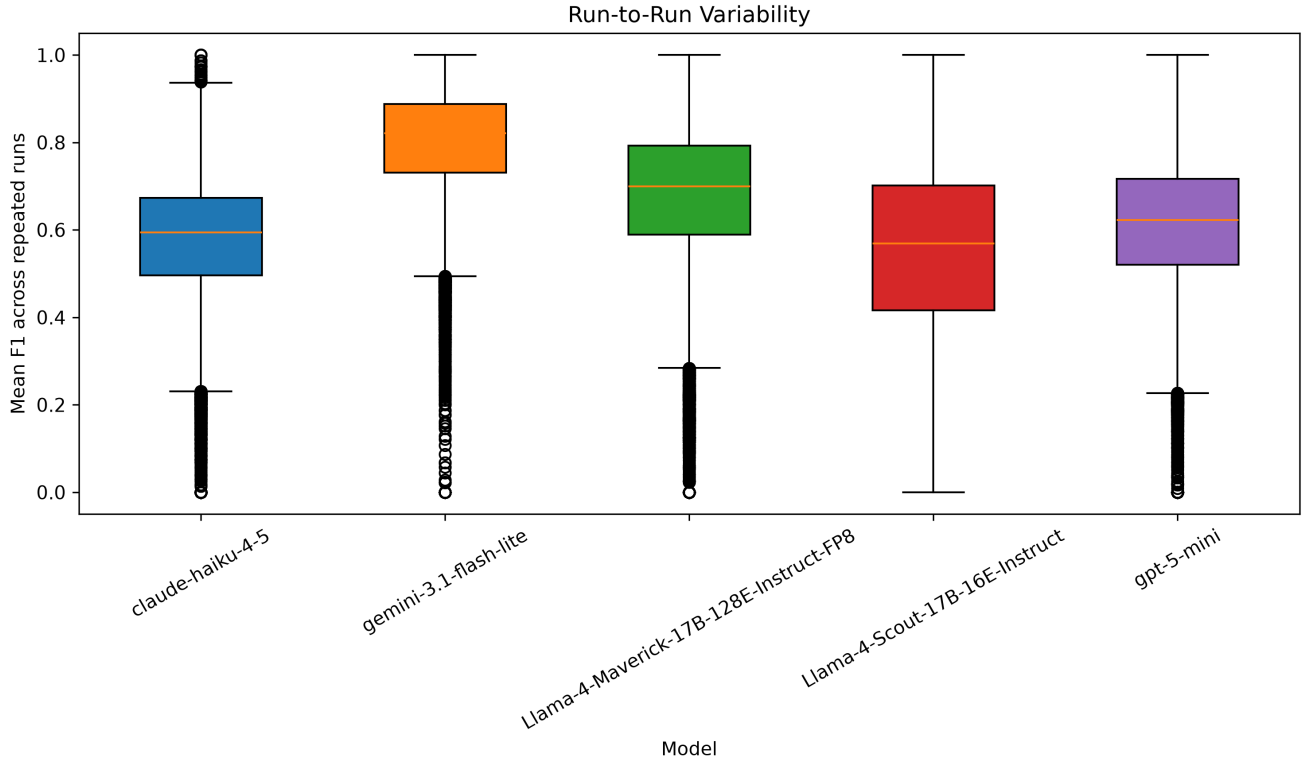


Figure 5: Run-to-Run Variability.

`gpt-5-mini` and `claude-haiku-4-5` show lower median values of approximately 0.62 and 0.59, respectively, whereas `Llama-4-Scout-17B-16E-Instruct` has the lowest median at approximately 0.57 and the widest overall whisker range, extending from 0.0 to 1.0. Notably, four models exhibit a dense concentration of lower-end outliers extending toward 0.0, while all five models contain instances reaching both very low and very high F1 values. This indicates that higher central retrieval performance does not eliminate substantial variation across individual instances. However, the plot alone cannot establish whether the observed variation reflects stochastic run-to-run differences, consistent failures on particular questions, or other experimental factors, since the metric aggregates repeated runs at the item level. Therefore, the figure provides descriptive evidence of variability but does not, by itself, establish the statistical significance or specific source of the observed differences.

Table 7 reports run-to-run dispersion metrics. Across all models, mean standard deviations strictly exceeded median standard deviations, indicating right-tailed stochastic instability. Run-to-run F1 ranges under identical prompts averaged between 0.141258 (gpt-5-mini) and 0.253241 (Llama-4-Scout). These stochastic run-to-run ranges exceeded aggregate demographic F1 gaps (0.0028 to 0.0258) by roughly an order of magnitude.

In 8,998 condition rows, models exhibited complete deterministic stability across repeated runs (SD = 0.0, range = 0.0). Of these rows, 2,846 represented complete deterministic retrieval failures where mean F1 was 0.0 across all five runs.

5.7. Error Profiles, Catastrophic Failure, and Lexical Consistency

Entity-level error composition across models, categories, and thresholds was characterized predominantly by omission rather than hallucination. Mean false negatives (5.016 to 5.066 entities per prompt across categories) consistently doubled mean false positives (2.338 to 2.367 entities).

Models displayed distinct generation volume profiles. Calibrated models tracked reference target entity volumes

Table 7
Stochastic run-to-run dispersion metrics across five repeated runs.

Model	Mean SD (F1)	Median SD (F1)	Mean F1 Range	Mean CV
Llama-4-Scout-17B-16E-Instruct	0.108343	0.089443	0.253241	0.277550
claude-haiku-4-5	0.080516	0.068313	0.188366	0.166988
Llama-4-Maverick-17B-128E-Instruct-FP8	0.070659	0.054772	0.165534	0.119653
gemini-3.1-flash-lite	0.081545	0.063246	0.190997	0.107765
gpt-5-mini	0.059296	0.044721	0.141258	0.117536

closely: gemini-3.1-flash-lite exhibited an entity deficit $<$ 0.17 (count ratio 0.987 to 0.996), Llama-4-Scout a deficit of -0.21 to -0.49 entities (ratio 0.941 to 0.975), and Llama-4-Maverick an entity surplus of $+1.12$ to $+1.16$ entities (ratio 1.28 to 1.29). In contrast, gpt-5-mini and claude-haiku-4-5 recorded severe omission profiles: gpt-5-mini had an entity count ratio of 0.558 with mean false negatives of 8.37, while claude-haiku-4-5 recorded a ratio of 0.578 with mean false negatives of 8.83. However, these omission metrics must be contextualized by benchmark ground-truth discrepancies: mean ground-truth targets for claude-haiku-4-5 (15.865) and gpt-5-mini (15.297) were substantially higher than those for gemini-3.1-flash-lite (9.384), Llama-4-Scout (8.518), and Llama-4-Maverick (7.948), confounding direct cross-model raw error count comparisons.

Decomposing complete retrieval failures (F1 $=$ 0.0) versus conditional retrieval quality (F1 $>$ 0) identified the primary mechanism underlying demographic differences in Llama-4-Scout. When Llama-4-Scout retrieved at least one matching entity, conditional retrieval quality (mean_f1_nonzero) remained uniform across demographic groups, spanning between 0.5768 and 0.5836 ($\Delta < 0.007$). In gemini-3.1-flash-lite, conditional F1 spanned from 0.7969 to 0.8001 ($\Delta < 0.004$). However, in Llama-4-Scout, complete retrieval failure rates (pct_f1_zero) showed marked disparities: asian (8.50%, n $=$ 2,218), hispanic (8.34%, n $=$ 2,177), and arab (7.70%, n $=$ 2,010) fact patterns experienced elevated failure rates relative to white (4.75%, n $=$ 1,239), homeless (4.81%, n $=$ 1,256), and wealthy (4.98%, n $=$ 1,300) prompts (total zero-retrieval instances $=$ 16,061 / 261,000). Total zero-retrieval counts for other models were substantially lower: 202 / 261,250 for gemini (0.0038% to 0.1608%), 888 / 261,250 for gpt-5-mini (0.30% to 0.38%), 1,620 / 261,250 for Maverick (0.58% to 0.68%), and 8,659 / 261,250 for claude-haiku (3.28% to 3.43%).

To evaluate lexical rule formulation stability, pairwise Jaccard similarities were computed between retrieved entity string sets across demographic variants:

$$\text{Jaccard}(S_i, S_j) = \frac{|S_i \cap S_j|}{|S_i \cup S_j|} \tag{21}$$

Table 8 reports counterfactual lexical consistency across models. These Jaccard metrics remained invariant across Hungarian matching thresholds ($\tau = 0.70$ to 0.90), as lexical sets were evaluated prior to semantic matching.

gpt-5-mini exhibited near-complete turnover of retrieved entity text strings across demographic prompts, recording a mean pairwise Jaccard similarity of 0.022160 (SD $=$ 0.016598), despite maintaining a macro semantic F1 of 0.608248. Demographic substitutions in gpt-5-mini induced extensive surface rephrasing, token rearrangement, and synonym substitution without destroying semantic retrieval.

5.8. Statistical Significance and Effect Magnitude

In non-parametric paired testing across 1,045 questions pooled across all models and thresholds, Wilcoxon signed-rank tests reached statistical significance ($p < 0.05$) in 35 of 45 pairwise demographic category comparisons. Extreme significance appeared in contrasts such as hispanic versus wealthy (p $=$ 3.25e-08), asian versus wealthy (p $=$ 3.45e-08), and female versus hispanic (p $=$ 5.31e-05). Conversely, 10 pairwise comparisons were non-significant, including female versus low-income ($p = 0.957$), homeless versus wealthy (p $=$ 0.361), and black versus white (p $=$ 0.314).

Standardized paired effect sizes (Cohen's $d_z = \frac{\text{mean}(D)}{\text{sd}(D)}$) across all 45 pairwise contrasts ranged from -0.205783 to $+0.169607$. Of the 45 contrasts, 27 exhibited negligible effect sizes ($|d_z| < 0.10$), 17 exhibited small effect sizes ($0.10 \leq |d_z| < 0.20$), and exactly 1 contrast fell into the small-to-medium range ($|d_z| \geq 0.20$): hispanic versus wealthy (d_z $=$ -0.205783). The largest positive effect size was female versus hispanic ($d_z = 0.169607$), and the smallest was female versus low-income ($d_z = 0.009599$). High statistical significance in Wilcoxon signed-rank tests was driven by the large sample size (n $=$ 1,045) detecting substantively negligible differences spanning 0.00027 to 0.00664 F1 points.

5.9. Empirical Nuances and Data Anomalies

The empirical findings present three core nuances:

1. *Macro Parity versus Intra-Question Instability*: Linear mixed-effects regression showed null demographic main effects across the population ($p \geq 0.506593$), yet 99.21% of question-condition profiles exhibited non-zero dispersion across demographic variants, with individual question ranges reaching 0.8795 in Llama-4-Scout and 1.0 in gemini-3.1-flash-lite. Aggregation masking explains this divergence: because demographic perturbations induce both positive and negative shifts across individual queries, item-level volatility cancels out in corpus-level aggregates.

Table 8
Pairwise lexical Jaccard similarity across demographic variants.

Model	Mean Pairwise Jaccard	SD (Jaccard)
Llama-4-Maverick-17B-128E-Instruct-FP8	0.581028	0.106861
Llama-4-Scout-17B-16E-Instruct	0.542061	0.110137
gemini-3.1-flash-lite	0.525312	0.093949
claude-haiku-4-5	0.455038	0.108240
gpt-5-mini	0.022160	0.016598

2. *Statistical Detectability versus Practical Substance*: Wilcoxon signed-rank tests indicated highly significant demographic differences ($p = 3.25\text{e-}08$), whereas standardized effect sizes remained uniformly small or negligible ($|d_z| \leq 0.205783$). High statistical power detected minor numerical differences lacking practical retrieval significance.
3. *Lexical Turnover versus Semantic Alignment*: In gpt-5-mini, counterfactual lexical Jaccard similarity dropped to 0.022160, while macro semantic F1 remained stable at 0.608248. Demographic substitutions triggered extensive lexical rephrasing without degrading semantic alignment with target legal rules.

Three data anomalies qualify these observations:

- *Sample Size Deficit in Llama-4-Scout*: Llama-4-Scout evaluated 260,950 total rows (or 261,000 cases), exactly 50 rows short per threshold (52,150 vs. 52,200) and 250 rows short overall compared to other models (261,250). In Friedman tests, Scout evaluated $n = 1,044$ questions versus 1,045 for others due to dropped queries or runtime timeouts specifically on `mbe_7`.

- *Ground-Truth Entity Count Discrepancies*: Mean ground-truth entity counts diverged substantially across models on the same benchmark: 15.865 entities for claude-haiku-4-5 (mean TP = 7.039, FN = 8.826) and 15.297 for gpt-5-mini (mean TP = 6.924, FN = 8.373), compared to 9.384 for gemini-3.1-flash-lite (TP = 7.457, FN = 1.927), 8.518 for Llama-4-Scout (TP = 4.808, FN = 3.710), and 7.948 for Llama-4-Maverick (TP = 5.568, FN = 2.380). Benchmark target sets were dynamically conditioned or filtered during extraction, confounding direct cross-model raw error count comparisons.

- *Threshold Non-Monotonicity*: Exactly 129 complete question-variant profiles in gemini-3.1-flash-lite exhibited non-monotonic threshold behavior, where calculated F1 increased as the Hungarian threshold rose from 0.70 to 0.90 due to global bipartite edge reassignments under stricter stringency criteria.

5.10. Discussion of the Research Hypothesis

The research hypothesis posited that introducing legally irrelevant demographic and social descriptors into identical legal fact patterns significantly alters retrieved legal rules, reducing retrieval consistency and stability. This hypothesis is partially supported with critical nuances:

- *Supported Aspects*: The stability degradation component of the hypothesis was confirmed at the intra-question counterfactual level. Swapping demographic descriptors altered retrieved rule formulations and induced within-question F1 dispersion across 99.21% of question-condition profiles (mean ranges 0.0894 to 0.2019; maximum ranges up to 1.0). In Llama-4-Scout, demographic perturbations produced statistically significant Friedman rank instability ($\chi^2 = 69.58$, $p = 1.84\text{e-}11$), significant mixed-effects interaction terms across 7 demographic categories, and an elevated complete retrieval failure rate for racial and ethnic minority categories (asian 8.50%, hispanic 8.34%, arab 7.70% vs. white 4.75%). In gpt-5-mini, demographic perturbations induced near-complete lexical turnover (Jaccard = 0.0222).

- *Unsupported Aspects*: The hypothesis that demographic perturbations cause systemic, population-level retrieval degradation or uniform directional bias was refuted. Across 1.3 million evaluations, mixed-effects demographic main effects were completely null (all $p \geq 0.506593$), aggregate cross-category F1 spreads were negligible (0.0067), effect sizes were small or negligible ($|d_z| \leq 0.205783$), and four of five models exhibited omnibus demographic rank invariance ($p > 0.05$). Stricter semantic thresholds did not interact with demographic categories (all 36 interaction terms $p \geq 0.5269$). Furthermore, stochastic run-to-run sampling variance exceeded cross-demographic variance by an order of magnitude.

5.11. Limitations and Interpretation Boundaries

Several methodological boundaries govern the interpretation of these findings:

1. *Causal Attribution*: While counterfactual prompt pairs isolate demographic descriptors as the trigger for intra-question dispersion, the evidence does not establish intentional or structural demographic prejudice against specific protected classes. Macro demographic parity and directional cross-model inversions indicate that localized retrieval instability is driven by prompt sensitivity and token shifting rather than uniform demographic bias.

2. *Multiple Testing and Aggregation in Paired Tests*: Pairwise Wilcoxon tests evaluated 45 contrasts without multiple-testing adjustments (e.g., Bonferroni-adjusted $\alpha = 0.00111$), elevating family-wise Type I error rates. Furthermore, these pairwise Wilcoxon tests and Cohen's d_z effect sizes pooled across all models and thresholds, masking model-specific nuances.
3. *Distributional Assumptions*: The linear mixed-effects model assumed Gaussian residuals on condition-averaged F1 scores bounded in $[0, 1]$, which exhibit point masses at zero due to complete retrieval failures.
4. *Lexical Metric Constraints*: Normalized string Jaccard similarity measures exact lexical token overlap and does not account for legal synonymy or doctrinal paraphrasing. Consequently, extreme lexical turnover in gpt-5-mini indicates surface rephrasing rather than definitive doctrinal failure.

6. Conclusion and Future Work

6.1. Conclusion

This study investigated whether legally irrelevant demographic and social descriptors can affect the legal rules retrieved by large language models (LLMs) from otherwise identical legal fact patterns. Unlike conventional legal benchmarks that primarily evaluate final-answer correctness, the study examined legal-rule retrieval as an intermediate component of legal reasoning. Using 1,045 Multistate Bar Examination (MBE) fact patterns, ten controlled demographic and socioeconomic variants, five LLMs, five semantic matching thresholds, and five repeated generations per condition, the evaluation was designed to distinguish counterfactual demographic sensitivity from ordinary generation-level variability.

The results provide evidence that demographic perturbations can be associated with substantial *within-question instability in legal-rule retrieval*. Across the experimental conditions, 99.21% of complete question-model-threshold profiles exhibited non-zero F1 dispersion across demographic variants. In some individual cases, the range between demographic variants was very large, reaching 1.0. The effect was particularly pronounced for Llama-4-Scout, for which demographic rank instability was statistically significant, while gpt-5-mini exhibited extremely low lexical Jaccard similarity across demographic variants despite maintaining comparatively stable semantic F1. These findings indicate that semantically similar or legally equivalent retrieval behaviour can sometimes be expressed through substantially different lexical formulations, while in other cases demographic perturbations can coincide with meaningful changes in the retrieved legal-rule set.

At the same time, the findings do not support the stronger interpretation that demographic descriptors produce a uniform, population-level degradation in legal-rule retrieval. Across the full evaluation corpus, demographic main effects in the mixed-effects model were not statistically significant, aggregate cross-category F1 differences were small, and four of the five evaluated models exhibited omnibus demographic rank invariance. Moreover, the direction of demographic variation differed across model architectures: a demographic category associated with lower retrieval performance for one model could be associated with higher performance for another. Thus, the observed phenomenon is better characterized as *localized counterfactual retrieval instability and architectural heterogeneity* than as a consistent directional penalty against a particular demographic group.

The study also demonstrates that demographic sensitivity cannot be interpreted independently of generation variability. Five repeated generations were therefore used for each experimental condition, and the results showed substantial run-to-run dispersion across the evaluated models. Indeed, the analysis indicates that stochastic sampling variance exceeded cross-demographic variance by an order of magnitude. This finding reinforces the importance of repeated evaluation when testing counterfactual sensitivity: a difference between two individual outputs should not automatically be attributed to the demographic perturbation when the same model and fact pattern can itself produce substantially different retrieval behaviour across repeated generations.

Semantic matching thresholds also materially affected measured retrieval performance. Increasing the Hungarian matching threshold from 0.70 to 0.90 produced significant negative effects on F1, while the demographic-by-threshold interactions were not statistically significant. This suggests that stricter semantic matching primarily changed the overall retrieval criterion rather than systematically amplifying or reducing demographic differences. The threshold analysis therefore further demonstrates that conclusions about legal-rule retrieval sensitivity depend partly on how semantic equivalence between retrieved rules and ground-truth rules is operationalized.

Taken together, these findings support the use of *counterfactual stability of legal-rule retrieval* as a complementary evaluation dimension for legal LLMs. A model can maintain acceptable aggregate retrieval performance while exhibiting substantial variation on individual fact patterns when legally irrelevant information is changed. Consequently, conventional aggregate accuracy or F1 alone may not fully characterize the stability of the legal knowledge selected by an LLM. The proposed evaluation framework provides a way to expose this distinction by holding legally relevant facts constant and testing whether the intermediate legal knowledge identified by the model remains stable under controlled perturbations.

Importantly, the findings should not be interpreted as evidence of intentional or structural demographic prejudice by the evaluated models. The study's counterfactual design establishes that changing demographic descriptors can coincide with changes in retrieval behaviour, but the observed macro-level parity, cross-model directional inversions, and

substantial stochastic variability limit stronger causal interpretations about discriminatory intent or systematic group-based disadvantage.

6.2. Limitations

Several limitations define the interpretation of the present results. First, the study uses MBE fact patterns and therefore evaluates legal-rule retrieval within a standardized examination setting. Although MBE questions provide controlled and legally grounded fact patterns, they do not represent the full diversity of real-world legal cases, jurisdictions, procedural contexts, or legal documents encountered in professional practice. The observed behaviour should therefore be validated on broader legal corpora and tasks before being generalized to deployed legal systems.

Second, the statistical analysis has methodological constraints. The paired Wilcoxon analyses included 45 demographic contrasts without a multiple-testing correction, which increases the possibility of family-wise Type I error. In addition, the pairwise effect sizes were pooled across models and thresholds, potentially masking model-specific behaviour. The linear mixed-effects model also assumes Gaussian residuals for condition-level F1 values bounded between zero and one, including point masses generated by complete retrieval failures.

Third, the semantic evaluation itself imposes limitations. The retrieval metric depends on embedding-based semantic matching and a Hungarian bipartite assignment procedure, while the lexical Jaccard analysis measures exact token overlap. In particular, lexical similarity does not capture legal synonymy, doctrinal equivalence, or legitimate paraphrasing. Consequently, substantial lexical turnover, such as that observed for gpt-5-mini, cannot by itself be interpreted as evidence that the model retrieved different legal doctrines.

Finally, the results demonstrate that demographic perturbations, stochastic generation, model architecture, semantic thresholds, and question-specific characteristics interact in complex ways. The study therefore establishes a measurable phenomenon of retrieval instability but does not identify a single underlying mechanism responsible for every observed change. In particular, the evidence does not establish that the observed instability originates exclusively from demographic representations, rather than from interactions between demographic tokens, prompt structure, stochastic sampling, and model-specific retrieval behaviour.

6.3. Future Work

Future research can extend this work in several directions. First, the evaluation should be expanded beyond MBE questions to additional legal benchmarks, jurisdictions, statutory materials, case law, and realistic legal research tasks. Such an expansion would determine whether counterfactual legal-rule instability generalizes beyond standardized examination fact patterns and whether the magnitude and structure of the effect differ across areas of law.

Second, future studies should increase the number of repeated generations and explicitly model the variance attributable to stochastic sampling. The present results demonstrate that five repetitions can reveal substantial run-to-run dispersion; larger repetition counts could provide more precise estimates of within-condition variability and improve separation between stochastic variation and systematic counterfactual effects. This would also permit more detailed variance decomposition across models, prompts, demographic conditions, and individual legal questions.

Third, future work should improve the semantic evaluation of retrieved legal rules. Human expert annotation could be combined with embedding-based matching to distinguish genuine doctrinal differences from alternative formulations of the same legal principle. Multiple semantic matching methods, legal-domain encoders, and expert-validated equivalence classes could reduce the dependence of conclusions on a particular similarity threshold. This is particularly important because lexical turnover may reflect paraphrasing rather than substantive retrieval failure.

Fourth, future research should investigate the mechanisms underlying the observed counterfactual instability. Controlled experiments could vary the position, wording, prominence, and frequency of demographic descriptors while keeping their semantic content constant. Such experiments could help determine whether retrieval changes arise primarily from lexical associations, contextual attention, prompt-position effects, or interactions between demographic descriptors and particular legal concepts. Comparing models with different architectures and training procedures may further clarify why the magnitude and direction of demographic sensitivity differ across systems.

Fifth, the proposed evaluation framework can be extended from *measurement to mitigation*. Once reliable sources of retrieval instability have been identified, future studies could evaluate interventions such as counterfactual training, retrieval re-ranking, fairness-aware retrieval, representation-level interventions, or structured legal knowledge graphs. Such methods should be evaluated not only for whether they reduce demographic sensitivity, but also for whether they preserve legal-rule recall, precision, semantic fidelity, and overall legal reasoning performance.

Finally, future work should examine whether instability at the legal-rule retrieval stage propagates to downstream reasoning and final legal conclusions. The present study deliberately isolates retrieval as an intermediate variable. An important next step is therefore to construct an end-to-end evaluation in which retrieved rules are passed to the subsequent reasoning stage and the effect of counterfactual retrieval changes on legal conclusions, explanations, and citations is measured. This would establish whether the retrieval instability identified here remains an intermediate phenomenon or translates into consequential differences in downstream legal reasoning.

Overall, this study suggests that evaluating legal LLMs requires attention not only to whether a model produces a correct final answer, but also to whether the legal knowledge

it retrieves remains stable when legally irrelevant aspects of a fact pattern are changed. Counterfactual legal-rule retrieval provides a complementary perspective for auditing this property and offers a foundation for future research on more stable, reproducible, and legally grounded language-model reasoning.

References

[1] Jeremias Adams-Prassl, Reuben Binns, and Aislinn Kelly-Lyth. Directly Discriminatory Algorithms. *The Modern Law Review*, 86(1):144–175, January 2023.

[2] Xuechunzi Bai, Angelina Wang, Ilia Sucholutsky, and Thomas L. Griffiths. Explicitly unbiased large language models still form biased associations. *Proceedings of the National Academy of Sciences*, 122(8):e2416228122, February 2025.

[3] Tilman Beck, Hendrik Schuff, Anne Lauscher, and Iryna Gurevych. Deconstructing the Effect of Sociodemographic Prompting.

[4] Hua Cai, Shuang Zhao, Liang Zhang, Xuli Shen, Qing Xu, Weilin Shen, Zihao Wen, and Tianke Ban. Unilaw-R1: A Large Language Model for Legal Reasoning with Reinforcement Learning and Iterative Inference, October 2025.

[5] Ke Chen, Yufei Zhou, Xitong Zhang, and Haohan Wang. Prompt Stability Matters: Evaluating and Optimizing Auto-Generated Prompt in General-Purpose Systems, May 2025. arXiv:2505.13546 [cs.AI].

[6] Hao-Fei Cheng, Logan Stapleton, Anna Kawakami, Venkatesh Sivaraman, Yanghuidi Cheng, Diana Qing, Adam Perer, Kenneth Holstein, Zhiwei Steven Wu, and Haiyi Zhu. How Child Welfare Workers Reduce Racial Disparities in Algorithmic Decisions. In *CHI Conference on Human Factors in Computing Systems*, pages 1–22, New Orleans LA USA, April 2022. ACM.

[7] Odysseas S. Chlapanis, Dimitrios Galanis, Nikolaos Aletras, and Ion Androutsopoulos. GreekBarBench: A Challenging Benchmark for Free-Text Legal Reasoning and Citations. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng, editors, *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 25099–25119, Suzhou, China, November 2025. Association for Computational Linguistics.

[8] Zhibo Chu, Zichong Wang, and Wenbin Zhang. Fairness in Large Language Models: A Taxonomic Survey. *ACM SIGKDD Explorations Newsletter*, 26(1):34–48, July 2024.

[9] Daniel S Cromvelle. ALGORITHMIC DISCRIMINATION AND EQUAL PROTECTION LAW: LEGAL REMEDIES FOR AI BIAS IN AUTOMATED DECISION- MAKING.

[10] Jiaxin Cui and Rohan Alexander. Same Prompt, Different Outcomes: Evaluating the Reproducibility of Data Analysis by LLMs, February 2026. arXiv:2602.14349 [stat.AP].

[11] Yongfu Dai, Duanyu Feng, Jimin Huang, Haochen Jia, Qianqian Xie, Yifang Zhang, Weiguang Han, Wei Tian, and Hao Wang. LAiW: A Chinese Legal Large Language Models Benchmark.

[12] J. De Curtò and I. De Zarzà. Metamorphic Testing for Semantic Invariance in Large Language Models. *IEEE Access*, 13:214772–214791, 2025.

[13] Fatemeh Dehghani, Roya Dehghani, Yazdan Naderzadeh Ardebili, and Shahryar Rahnamayan. Large Language Models in Legal Systems: A Survey. *Humanities and Social Sciences Communications*, 12(1):1977, December 2025.

[14] Ameet Deshpande, Vishvak Murahari, Tanmay Rajpurohit, Ashwin Kalyan, and Karthik Narasimhan. Toxicity in chatgpt: Analyzing persona-assigned language models. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1236–1270, Singapore, December 2023. Association for Computational Linguistics.

[15] David Esiobu, Xiaoqing Tan, Saghar Hosseini, Megan Ung, Yuchen Zhang, Jude Fernandes, Jane Dwivedi-Yu, Eleonora Presani, Adina Williams, and Eric Smith. ROBBIE: Robust Bias Evaluation of Large Generative Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3764–3814, Singapore, 2023. Association for Computational Linguistics.

[16] Yu Fan, Jingwei Ni, Jakob Merane, Yang Tian, Yoan Hermstrüwer, Yinya Huang, Mubashara Akhtar, Etienne Salimbeni, Florian Geering, Oliver Dreyer, Daniel Brunner, Markus Leippold, Mrinmaya Sachan, Alexander Stremitzer, Christoph Engel, Elliott Ash, and Joel Niklaus. LEXAM:BENCHMARKINGLEGALREASONING ON340 LAWEXAMS. 2026.

[17] Xiao Fang, Shangkun Che, Minjia Mao, Hongzhe Zhang, Ming Zhao, and Xiaohang Zhao. Bias of AI-generated content: an examination of news produced by large language models. *Scientific Reports*, 14(1):5224, March 2024.

[18] Zhiwei Fei, Xiaoyu Shen, Dawei Zhu, Fengzhe Zhou, Zhuo Han, Alan Huang, Songyang Zhang, Kai Chen, Zhixin Yin, Zongwen Shen, Jidong Ge, and Vincent Ng. LawBench: Benchmarking Legal Knowledge of Large Language Models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7933–7962, Miami, Florida,

USA, 2024. Association for Computational Linguistics.

[19] Yi Feng, Chuanyi Li, and Vincent Ng. Legal Judgment Prediction: A Reflection on the State of the Art. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens, editors, *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 45254–45273, San Diego, California, United States, July 2026. Association for Computational Linguistics.

[20] Madeleine Forster. Refugee protection in the artificial intelligence era: a test case for rights. Technical report, Royal Institute of International Affairs, September 2022.

[21] Shyamrao A. Gade and Sivaram Ponnusamy. Statistical Analytical Review of Language Processing Models for Legal Outcome Predictions: Performance, Fairness, and Reasoning Fidelity Across Contemporary Frameworks. *International Journal of Computer Information Systems and Industrial Management Applications*, 18(9s):1348–1364, July 2026.

[22] Oren Gazal Ayal, Zohar Elyoseph, and Adir Solomon. Evaluating Large Language Models as Judicial Decision-Makers. *Justice Quarterly*, 43(5):1184–1219, July 2026.

[23] Talia B. Gillis. The Input Fallacy. *Minnesota Law Review*, 106(3):1175, 2022.

[24] Miguel Angel Alvarado Gonzalez, Michelle Bruno Hernandez, Miguel Angel Peñaloza Perez, Bruno Lopez Orozco, Jesus Tadeo Cruz Soto, and Sandra Malagon. Do Repetitions Matter? Strengthening Reliability in LLM Evaluations, September 2025. arXiv:2509.24086 [cs.AI].

[25] Faith Gordon. Virginia Eubanks (2018) Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor. New York: Picador, St Martin's Press. *Law, Technology and Humans*, 1:162–164, November 2019.

[26] Neel Guha, Julian Nyarko, Daniel E. Ho, Christopher Ré, Adam Chilton, Aditya Narayana, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel Rockmore, Diego Zambrano, Dmitry Talisman, Enam Hoque, Faiz Surani, Frank Fagan, Galit Sarfaty, Gregory M. Dickinson, Haggai Porat, Jason Hegland, Jessica Wu, Joe Nudell, Joel Niklaus, John Nay, Jonathan H. Choi, Kevin Tobia, Margaret Hagan, Megan Ma, Michael A. Livermore, Nikon Rasumov-Rahe, Nils Holzenberger, Noam Kolt, Peter Henderson, Sean Rehaag, Sharad Goel, Shang Gao, Spencer Williams, Sunny Gandhi, Tom Zur, Varun Iyer, and Zehua Li. Legalbench: A Collaboratively Built Benchmark for Measuring Legal Reasoning in Large Language Models, 2023.

[27] Jennifer Haase, Jana Gonnermann-Müller, Paul H. P. Hanel, Nicolas Leins, Thomas Kosch, Jan Mendling, and Sebastian Pokutta. Within-Model vs Between-Prompt Variability in Large Language Models for Creative Tasks, January 2026. arXiv:2601.21339 [cs.AI].

[28] Philipp Hacker. Teaching fairness to artificial intelligence: Existing and novel strategies against algorithmic discrimination under EU law. *Common Market Law Review*, 55(Issue 4):1143–1185, August 2018.

[29] John J Hanna, Abdi D Wakene, Andrew O Johnson, Christoph U Lehmann, and Richard J Medford. Assessing Racial and Ethnic Bias in Text Generation by Large Language Models for Health Care–Related Tasks: Cross-Sectional Study. *Journal of Medical Internet Research*, 27:e57257, March 2025.

[30] Deborah Hellman. MEASURING ALGORITHMIC FAIRNESS. 106(4):811–866, 2020.

[31] Tiancheng Hu, Yara Kyrychenko, Steve Rathje, Nigel Collier, Sander van der Linden, and Jon Roozenbeek. Generative language models exhibit social identity biases. *Nature Computational Science*, 5(1):65–75, January 2025.

[32] Cong Jiang and Xiaolei Yang. Legal Syllogism Prompting: Teaching Large Language Models for Legal Judgment Prediction. In *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law*, pages 417–421, Braga Portugal, June 2023. ACM.

[33] Daniel Martin Katz, Michael James Bommarito, Shang Gao, and Pablo Arredondo. GPT-4 Passes the Bar Exam, 2023.

[34] Joshua Kelsall, Xingwei Tan, Aislinn Bergin, Jiahong Chen, Maria Waheed, Tom Sorell, Rob Procter, Maria Liakata, Jenny Chim, and Serene Chi. A rapid evidence review of evaluation techniques for large language models in legal use cases: trends, gaps, and recommendations for future research. *AI & SOCIETY*, 41(4):4025–4043, April 2026.

[35] Hadas Kotek, Rikker Dockum, and David Sun. Gender bias and stereotypes in Large Language Models. In *Proceedings of The ACM Collective Intelligence Conference*, CI '23, pages 12–24, New York, NY, USA, November 2023. Association for Computing Machinery.

[36] Jiatong Li, Renjun Hu, Kunzhe Huang, Yan Zhuang, Qi Liu, Mengxiao Zhu, Xing Shi, and Wei Lin. PertEval: Unveiling Real Knowledge Capacity of LLMs with Knowledge-Invariant Perturbations.

[37] Paul Pu Liang, Chiyu Wu, Louis-Philippe Morency, and Ruslan Salakhutdinov. Towards Understanding and Mitigating Social Biases in Language Models.

In *Proceedings of the 38th International Conference on Machine Learning*, pages 6565–6576. PMLR, July 2021.

[38] Antoine Louis, Gijs Van Dijck, and Gerasimos Spanakis. Interpretable Long-Form Legal Question Answering with Retrieval-Augmented Large Language Models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(20):22266–22275, March 2024.

[39] Varun Magesh, Faiz Surani, Matthew Dahl, Mirac Suzgun, Christopher D. Manning, and Daniel E. Ho. Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools. *Journal of Empirical Legal Studies*, 22(2):216–242, June 2025.

[40] Eric Martínez. Re-evaluating GPT-4's bar exam performance. *Artificial Intelligence and Law*, 33(3):581–604, September 2025.

[41] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A Survey on Bias and Fairness in Machine Learning. *ACM Computing Surveys*, 54(6):1–35, July 2022.

[42] Xin Miao, Yongqi Li, Hankun Kang, Mayi Xu, Jintao Wen, Yuyang Ren, and Tieyun Qian. How Robust are Large Language Models Against Word-Level Spurious Correlations? A Causal Discovery Approach. *Machine Learning*, 115(3):59, March 2026.

[43] Venkatesh Mishra, Bimsara Pathiraja, Mihir Parmar, Sat Chidananda, Jayanth Srinivasa, Gaowen Liu, Ali Payani, and Chitta Baral. Investigating the Shortcomings of LLMs in Step-by-Step Legal Reasoning.

[44] Moin Nadeem, Anna Bethke, and Siva Reddy. StereoSet: Measuring stereotypical bias in pretrained language models. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5356–5371, Online, August 2021. Association for Computational Linguistics.

[45] Roberto Navigli, Simone Conia, and Björn Ross. Biases in Large Language Models: Origins, Inventory, and Discussion. *Journal of Data and Information Quality*, 15(2):1–21, June 2023.

[46] John J Nay, David Karamardian, Sarah B Lawsky, Wenting Tao, Meghana Bhat, Raghav Jain, Travis Lee, Jonathan H Choi, and Jungo Kasai. Large language models as tax attorneys: a case study in legal capabilities emergence.

[47] Shubham Kumar Nigam, Aniket Deroy, Subhankar Maity, and Arnab Bhattacharya. Rethinking Legal Judgement Prediction in a Realistic Scenario in the Era of Large Language Models. In *Proceedings of the Natural Legal Language Processing Workshop 2024*, pages 61–80, Miami, FL, USA, 2024. Association for Computational Linguistics.

[48] Mahmud Omar, Shelly Soffer, Reem Agbareia, Nicola Luigi Bragazzi, Donald U. Apakama, Carol R Horowitz, Alexander W Charney, Robert Freeman, Benjamin Kummer, Benjamin S Glicksberg, Girish N Nadkarni, and Eyal Klang. Socio-Demographic Biases in Medical Decision-Making by Large Language Models: A Large-Scale Multi-Model Analysis, October 2024.

[49] Mahmud Omar, Vera Sorin, Reem Agbareia, Donald U. Apakama, Ali Soroush, Ankit Sakhuja, Robert Freeman, Carol R. Horowitz, Lynne D. Richardson, Girish N. Nadkarni, and Eyal Klang. Evaluating and addressing demographic disparities in medical large language models: a systematic review. *International Journal for Equity in Health*, 24(1):57, February 2025.

[50] Derya Ozkul and Refugee Studies Centre. Automating Immigration and Asylum.

[51] Paola Panarese, Marta Margherita Grasso, and Claudia Solinas. Algorithmic bias, fairness, and inclusivity: a multilevel framework for justice-oriented AI. *AI & SOCIETY*, 41(4):2803–2825, April 2026.

[52] Shounak Paul, Rajas Bhatt, Pawan Goyal, and Saptarshi Ghosh. Legal Statute Identification: A Case Study using State-of-the-Art Datasets and Methods. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2231–2240, Washington DC USA, July 2024. ACM.

[53] Gaojun Peng, Dapeng Li, Gang Wang, Shuang Guo, Lina Su, and Tianxue Yu. A review of the applications and tasks of large language models in the legal field. *PeerJ Computer Science*, 12:e3836, May 2026.

[54] Eric A. Posner and Shivam Saran. Judge AI: Assessing Large Language Models in Judicial Decision-Making, 2025.

[55] Philippe Prince Tritto and Hiram Ponce. Assessing AI-Generated Legal Reasoning: A Benchmark for Legal Text Quality from Literature Review. In Lourdes Martínez-Villaseñor, Bella Martínez-Seis, and Obdulia Pichardo, editors, *Artificial Intelligence – COMIA 2025*, pages 54–68, Cham, 2025. Springer Nature Switzerland.

[56] Rebecca Qian, Candace Ross, Jude Fernandes, Eric Michael Smith, Douwe Kiela, and Adina Williams. Perturbation Augmentation for Fairer NLP. In *Proceedings of the 2022 Conference on Empirical*

Methods in Natural Language Processing, pages 9496–9521, Abu Dhabi, United Arab Emirates, 2022. Association for Computational Linguistics.

[57] Chantal Shaib, Vinith M Suriyakumar, Byron C Wallace, and Marzyeh Ghassemi. Learning the Wrong Lessons: Syntactic-Domain Spurious Correlations in Language Models.

[58] Emily Sheng, Kai-Wei Chang, Prem Natarajan, and Nanyun Peng. Societal Biases in Language Generation: Progress and Challenges. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4275–4293, Online, 2021. Association for Computational Linguistics.

[59] Dong Shu, Haoran Zhao, Xukun Liu, David Demeter, Mengnan Du, and Yongfeng Zhang. LawLLM: Law Large Language Model for the US Legal System. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 4882–4889, Boise ID USA, October 2024. ACM.

[60] Ruihao Shui, Yixin Cao, Xiang Wang, and Tat-Seng Chua. A Comprehensive Evaluation of Large Language Models on Legal Judgment Prediction. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7337–7348, Singapore, 2023. Association for Computational Linguistics.

[61] Vera Sorin, Panagiotis Korfiatis, Jeremy D. Collins, Donald Apakama, Mahmud Omar, Benjamin S. Glicksberg, Mei-Ean Yeow, Megan Brandeland, Girish N. Nadkarni, and Eyal Klang. Socio-Demographic Modifiers Shape Large Language Models' Ethical Decisions. *Journal of Healthcare Informatics Research*, 9(4):567–586, December 2025.

[62] Moritz Staudinger, Wojciech Kusa, Florina Piroi, Aldo Lipani, and Allan Hanbury. A Reproducibility and Generalizability Study of Large Language Models for Query Generation. In *Proceedings of the 2024 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*, pages 186–196, Tokyo Japan, December 2024. ACM.

[63] Matthew Stubenberg, Chloe Berridge, Thomas Smith, and Joshua Casey. How AI Stacks Up Against the Multistate Bar Exam, May 2025.

[64] Zechen Sun, Yisheng Xiao, Juntao Li, Yixin Ji, Wenliang Chen, and Min Zhang. Exploring and Mitigating Shortcut Learning for Generative Large Language Models.

[65] Ruixiang Tang, Dehan Kong, Longtao Huang, and Hui Xue. Large Language Models Can be Lazy Learners: Analyze Shortcuts in In-Context Learning. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4645–4657, Toronto, Canada, 2023. Association for Computational Linguistics.

[66] Li Wang, Xi Chen, XiangWen Deng, Hao Wen, MingKe You, WeiZhi Liu, Qi Li, and Jian Li. Prompt engineering in consistency and reliability with the evidence-based guideline for LLMs. *npj Digital Medicine*, 7(1):41, February 2024.

[67] Tianlu Wang, Rohit Sridhar, Diyi Yang, and Xuezhi Wang. Identifying and Mitigating Spurious Correlations for Improving Robustness in NLP Models. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1719–1729, Seattle, United States, 2022. Association for Computational Linguistics.

[68] Xukang Wang, Ying Cheng Wu, Xueliang Ji, and Hongpeng Fu. Algorithmic discrimination: examining its types and regulatory measures with emphasis on US legal practices. *Frontiers in Artificial Intelligence*, 7:1320277, May 2024.

[69] Xuran Wang, Xinguang Zhang, Vanessa Hoo, Zhouhang Shao, and Xuguang Zhang. LegalReasoner: A Multi-Stage Framework for Legal Judgment Prediction via Large Language Models and Knowledge Integration. *IEEE Access*, 12:166843–166854, 2024.

[70] Yanbo Wang, Zixiang Xu, Yue Huang, Chujie Gao, Siyuan Wu, Jiayi Ye, Pin-Yu Chen, Xiuying Chen, and Xiangliang Zhang. Adaptive Distraction: Probing LLM Contextual Robustness with Automated Tree Search.

[71] Hilde Weerts, Raphaële Xenidis, Fabien Tarissan, Henrik Palmer Olsen, and Mykola Pechenizkiy. Algorithmic Unfairness through the Lens of EU Non-Discrimination Law: Or Why the Law is not a Decision Tree. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, pages 805–816, New York, NY, USA, June 2023. Association for Computing Machinery.

[72] Raphaële Xenidis. Tuning EU equality law to algorithmic discrimination: Three pathways to resilience. *Maastricht Journal of European and Comparative Law*, 27(6):736–758, December 2020.

[73] Alice Xiang. Reconciling Legal and Technical Approaches to Algorithmic Bias. *Tennessee Law Review*, 88(3):649, January 2021.

[74] Yifan Yang, Xiaoyu Liu, Qiao Jin, Furong Huang, and Zhiyong Lu. Unmasking and quantifying racial bias of large language models in medical report generation. *Communications Medicine*, 4(1):176, September 2024.

[75] Yu Yuan, Lili Zhao, Kai Zhang, Guangting Zheng, and Qi Liu. Do LLMs Overcome Shortcut Learning? An Evaluation of Shortcut Challenges in Large Language Models.

[76] Duzhen Zhang, Zhong-Zhi Li, Ming-Liang Zhang, Jiaxin Zhang, Zengyan Liu, Yuxuan Yao, Haotian Xu, Junhao Zheng, Xiuyi Chen, Yingying Zhang, Fei Yin, Jiahua Dong, Zhijiang Guo, Le Song, and Cheng-Lin Liu. From System 1 to System 2: A Survey of Reasoning Large Language Models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(3):3335–3354, March 2026.

[77] Kepu Zhang. SyLeR: A Framework for Explicit Syllogistic Legal Reasoning in Large Language Models. 2025.

[78] Li Zhang, Matthias Grabmair, Morgan Gray, and Kevin Ashley. Thinking Longer, Not Always Smarter: Evaluating LLM Capabilities in Hierarchical Legal Reasoning. In *Proceedings of the Symposium on Computer Science and Law*, pages 71–85, Berkeley CA USA, March 2026. ACM.

[79] Miao Zhang, Kelly Chen, Mehrab Tanjim, and Rumi Chunara. Identity-Robust Language Model Generation via Content Integrity Preservation.

[80] Lucia Zheng, Neel Guha, Javokhir Arifov, Sarah Zhang, Michal Skreta, Christopher D. Manning, Peter Henderson, and Daniel E. Ho. A Reasoning-Focused Legal Retrieval Benchmark. In *Proceedings of the Symposium on Computer Science and Law on ZZZ*, pages 169–193, Munich Germany, March 2025. ACM.

[81] Yuqing Zhou, Ruixiang Tang, Ziyu Yao, and Ziwei Zhu. Navigating the Shortcut Maze: A Comprehensive Analysis of Shortcut Learning in Text Classification by Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 2586–2614, Miami, Florida, USA, 2024. Association for Computational Linguistics.

[82] Miri Zilka, Holli Sargeant, and Adrian Weller. Transparency, Governance and Regulation of Algorithmic Tools Deployed in the Criminal Justice System: a UK Case Study. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '22, pages 880–889, New York, NY, USA, July 2022. Association for Computing Machinery.