The Lawyer That Never Thinks: Consistency and Fairness as Keys to Reliable AI

Dana Alsagheer, Abdulrahman Kamal, Mohammad Kamal, Cosmo Yang Wu, Weidong Shi
University of Houston
dralsagh@CougarNet.UH.EDU

Abstract

Large Language Models (LLMs) are increasingly used in high-stakes domains like law and research, yet their inconsistencies and response instability raise concerns about trustworthiness. This study evaluates six leading LLMs—GPT-3.5, GPT-4, Claude, Gemini, Mistral, and LLaMA 2—on rationality, stability, and ethical fairness through reasoning tests, legal challenges, and bias-sensitive scenarios. Results reveal significant inconsistencies, highlighting trade-offs between model scale, architecture, and logical coherence. These findings underscore the risks of deploying LLMs in legal and policy settings, emphasizing the need for AI systems that prioritize transparency, fairness, and ethical robustness.

1 Introduction

Large Language Models (LLMs) are increasingly integrated into high-stakes domains such as law, governance, and research, where they help analyze legal texts, generate structured arguments, and synthesize regulatory information (Bommasani et al., 2021; OpenAI, 2023d; Das et al., 2024). Although models like GPT-4 and Claude achieve near-human performance in standardized legal assessments (e.g., MBE), their lack of consistency and susceptibility to response drift raise significant concerns about fairness, reliability, and ethical deployment (OpenAI, 2023d; Katsumi and Liu, 2023; Zhong et al., 2023).

Despite their strong performance on accuracy-based benchmarks, LLMs frequently provide contradictory responses to semantically equivalent prompts, leading to biased or unfair decision-making in legal and policy contexts (Bender et al., 2021; Schramowski et al., 2022). Unlike human experts, who adhere to stable reasoning frameworks, LLMs often fail to generalize consistently across similar cases, creating ethical risks in AI-assisted legal judgments, governance, and social applications (Bommarito and Katz, 2022; Choi et al., 2023; Livermore and Southall, 2023). A core ethical concern is the impact of response variability on fairness and accountability. Inconsistent AI-generated legal reasoning can result in unequal treatment of cases, where identical inputs yield divergent interpretations (Livermore and Southall, 2023; Rawal et al., 2023). This unpredictability undermines trust in AI-assisted legal frameworks, particularly in scenarios where model outputs influence high-stakes decisions. In professional applications, where LLMs are expected to provide stable and legally sound interpretations, inconsistencies reinforce existing biases or create arbitrary legal precedents (Weidinger et al., 2021; Hendrycks et al., 2021).

Beyond law, LLM inconsistencies extend to other critical domains. In medical AI, studies show that models provide different diagnostic recommendations for the same symptoms depending on phrasing variations, introducing unfair disparities in patient treatment (Wang et al., 2023; Gao et al., 2023). Similarly, in financial forecasting, minor changes in input formatting can yield significantly different risk assessments, posing ethical concerns in AI-driven lending and insurance policies (Weidinger et al., 2022; Avrahami et al., 2023).

This instability in AI decision-making is particularly concerning, as logical consistency is essential for fair and unbiased outcomes (Rawal et al., 2023). If an AI system's reasoning fluctuates under similar conditions, it risks reinforcing biases, leading to unreliable legal, medical, or financial determinations and undermining trust in AI-assisted systems (Zhong et al., 2023). Unlike humans, who maintain structured cognitive frameworks to ensure stable reasoning, LLMs lack inherent mechanisms to enforce consistency, resulting in arbitrary or unpredictable outcomes across repeated trials (Bommasani et al., 2021; Zhong et al., 2023).

Table 1 presents an example of inconsistency in model responses across multiple queries. It illustrates the variability observed when six large

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 9943–9954

Table 1: Diagnostic Base Rate Neglect Results Across Models

Scenario and Question	Imagine that AIDS occurs in five in every 1,000 people. Imagine also that there is a test to diagnose the disease that always gives a positive result when a person has AIDS. Finally, imagine that the test has a false positive rate of 15 percent. This means that the test wrongly indicates that AIDS is present in 15 percent of the cases where the person does not have AIDS. Imagine that we choose a person to administer the test to randomly, and that it yields a positive result (indicates that the person has AIDS). **Question:** What is the probability that the individual has AIDS, assuming that we know nothing else about the individual's personal or medical history? **Base Rate Neglect:** 85% **Correct Answer:** 3% (Burgoyne et al., 2023)					
Model	**ChatGPT-3.5**	**GPT-4**	**Claude**	**Llama 3.2**	**Gemini**	**Mistral**
First Attempt (%)	25.0	25.0	10.0	85.0	14.0	73.0
Second Attempt (%)	54.0	26.0	5.2	5.0	10.0	60.0

language models are given the same query multiple times. Despite an identical input, the models generate different outputs, exposing inconsistencies in reasoning and decision-making. This variability underscores the need for architectural refinements, fine-tuning strategies, and response filtering to improve stability, particularly in high-stakes applications where consistency is crucial for reliable AI deployment.

Existing AI evaluations primarily focus on bias detection in training data but fail to account for decision stability over multiple test trials (Bender et al., 2021; Stanovich, 2011). Unlike accuracy-based benchmarks, fairness-aware AI evaluation must consider logical stability across repeated interactions. If a model exhibits response drift, it can introduce unintended biases in legal and policy applications, disproportionately affecting marginalized communities (Hendrycks et al., 2021; Ferrara et al., 2023).

To address this issue, we assess LLM consistency by measuring response stability across multiple trials using a combination of statistical metrics, including the Test-Retest Consistency Score (TRCS), Intraclass Correlation Coefficient (ICC), and Analysis of Variance (ANOVA). This study introduces TRCS as a novel metric for evaluating LLM stability and fairness in high-stakes applications. Unlike traditional evaluations that prioritize accuracy alone, our approach examines multi-trial response patterns to identify inconsistencies that can lead to unreliable or biased decision-making.

Our study systematically analyzes consistency to understand LLM limitations better and offer insights into how AI can improve professional and legal decision-making.

1.1 Contributions

This study makes the following key contributions:

- **Consistency Evaluation Across LLMs:** Assessed response stability across six large language models (LLMs) with varying architectures and parameter sizes.
- **Novel Consistency Metric:** Introduced the *Test-Retest Consistency Score (TRCS)* alongside the *Intraclass Correlation Coefficient (ICC)* and *Analysis of Variance (ANOVA)* for systematic variability analysis.
- **Hybrid Evaluation Dataset:** Developed a benchmark combining general reasoning tasks and legal questions to assess logical coherence and domain-specific legal reasoning.
- **Holistic Evaluation Framework:** Emphasized both accuracy and consistency to support robust and ethically grounded AI decision-making.

2 Related Work

OpenAI highlighted GPT-4's "human-level performance on various professional and academic benchmarks"(OpenAI, 2023e), particularly emphasizing its performance on the Uniform Bar Examination. OpenAI prominently reported that GPT-4 scored in or around the "90th percentile" or "the top 10% of test-takers"(OpenAI, 2023e,c,b). However, studies such as (Martínez, 2024) have raised questions about these claims, emphasizing the importance of assessing the model's final scores and analyzing its responses and explanations to determine the depth of its legal reasoning and consistency.

The Multistate Bar Examination (MBE) is a standardized multiple-choice test designed to assess

core legal knowledge and reasoning skills across seven key areas: Civil Procedure, Constitutional Law, Contracts, Criminal Law and Procedure, Evidence, Real Property, and Torts (Jayakumar et al., 2023; Riebe, 2006; Goforth, 2015; Barbri, 2024). Developed and administered by the *National Conference of Bar Examiners (NCBE)*, the MBE serves as a critical component of the bar admission process in most U.S. jurisdictions and contributes significantly to the overall Uniform Bar Examination (UBE) score (Bommarito and Katz, 2022; Katsumi and Liu, 2023; Heidemann, 2020).

The exam consists of 200 multiple-choice questions, split into two three-hour sessions (Nystrom, 2013; Simkovic and McIntyre, 2015). The MBE is designed to evaluate factual legal knowledge and assess a candidate's ability to apply legal principles, analyze fact patterns, and differentiate between complex legal arguments. Its standardized nature ensures that jurisdictions can objectively compare candidates across different regions, making it a widely used benchmark for assessing legal proficiency(Curcio, 2002; Schwartz, 2007; Johnson and White, 2021).

Recent advancements in natural language processing (NLP) have led to the increasing use of LLMs for legal applications, including legal research, document review, and decision support. Several legal NLP benchmarks, such as CaseHOLD (Zhong et al., 2020) and SPoT (Chalkidis et al., 2021), evaluate models on tasks like case retrieval and legal question answering. While these benchmarks assess text comprehension and legal knowledge retrieval, they do not examine whether models reason consistently across different formulations of the same legal issue. This limitation is particularly concerning in high-stakes legal applications, where inconsistencies in AI-generated responses could lead to incorrect interpretations of legal principles.

Most research on LLM legal reasoning prioritizes accuracy while neglecting consistency. Studies evaluating LLMs on legal and rationality tasks assess single-instance performance, disregarding whether models produce stable responses to identical or slightly altered prompts. Santurkar et al.(Santurkar et al., 2023) found that despite achieving high accuracy in formal logic, causal reasoning, and probabilistic inference, LLMs often fail to maintain consistency across trials. Similarly, Binz et al.(Binz and Schulz, 2023) demonstrated that LLMs generate hallucinated justifications and overconfident responses, but did not examine whether these errors persist across multiple queries.

A recent study (Macmillan-Scott and Musolesi, 2024b) evaluated LLMs using cognitive psychology tasks to assess rationality, revealing that their irrationality manifests differently from human reasoning. While human errors tend to follow predictable cognitive biases, LLM responses are highly variable and inconsistent across repeated trials. This lack of stability raises concerns for applications in law and decision-making, where logical coherence is just as critical as accuracy.

Despite these insights, most legal and rationality benchmarks prioritize accuracy while overlooking consistency. For example, Zhong et al.(Zhong et al., 2020) assessed LLMs on legal entailment tasks. Still, their study only considered single-instance correctness, failing to capture whether models apply legal reasoning stably across multiple trials. Similarly, studies evaluating GPT-4's bar exam performance focus on accuracy metrics without investigating whether the model maintains consistent answers when given the exact legal prompt multiple times (Bommarito and Katz, 2022; Livermore and Southall, 2023; Choi et al., 2023). While prior research has advanced the evaluation of LLM rationality, most studies conflate accuracy with reasoning ability, neglecting a critical aspect of human decision-making: consistency. In psychology and cognitive science, rational decision-making is about correctness and the ability to apply stable reasoning across repeated or slightly modified scenarios (Stanovich, 2011).

However, existing LLM evaluations fail to measure consistency, leading to cases where models provide contradictory answers to the same legal question depending on minor rewording or repetition. This inconsistency highlights a fundamental limitation in the generalization of LLMs. A well-generalized model should perform well on unseen tasks and maintain logical coherence when faced with repeated questions. The inability to do so suggests that LLMs rely heavily on surface-level patterns rather than genuinely understanding and internalizing legal reasoning structures. This limitation is particularly critical in legal AI applications, where unpredictable model behavior could undermine trust in AI-assisted decision-making.

Table 2: Summary of Large Language Models (LLMs) by size, availability, and key features.

Model	Size (Params)	Availability	Key Features	Cite
Mistral	7B–12B	Free	Lightweight and efficient for smaller-scale tasks.	(AI, 2023c)
Claude 2	∼52–100B	Free/Paid	Safety-focused, developed by Anthropic.	(Anthropic, 2023)
LLaMA 3.2 (70B)	70B	Free	Open-source, widely used for research under Meta's license.	(AI, 2023a)
Gemini 1.5	∼70B	Free/Paid	High reasoning performance, competitive with GPT-4.	(DeepMind, 2023)
ChatGPT-3.5	∼175B	Free/Paid	Reliable and versatile for general use.	(OpenAI, 2023a)
GPT-4	1–1.76T	Paid	Excels in complex reasoning and problem-solving.	(OpenAI, 2023d)

3 Construction of Our Evaluation

Our methodology combines domain-specific legal evaluations with well-established rationality benchmarks to assess the reasoning capabilities of large language models (LLMs). We treat each model as an independent evaluation unit, enabling fair comparisons across architectures, training regimes, and parameter scales. To systematically assess reliability, we employ robust statistical measures, including the *Test-Retest Consistency Score (TRCS)*, *Intraclass Correlation Coefficient (ICC)*, and *Analysis of Variance (ANOVA)*.

We conducted six trials over 30 days. In each trial, we presented every question twice using identical prompts, resulting in twelve responses per question per model. This repeated-measures design controls for prompt variation and allows us to isolate internal variability in model reasoning across both short-term and medium-term timeframes. Rather than administering all trials consecutively, we deliberately distributed the testing schedule to capture temporal fluctuations in model behavior. This design is essential for evaluating high-stakes applications such as legal analysis and rational decision-making, where consistent and coherent reasoning is critical. Any inconsistencies observed across trials are attributed to internal instability—such as sampling randomness, latent uncertainty, or backend model updates—rather than variation in input structure.

All code, Python scripts, and data used in this study are publicly available at our GitHub repository: the *Rationality of Large Language Models* project.[1]

[1] https://github.com/hala00001/Rationality-of-Large-language-models-

3.1 Evaluation Framework

We evaluated six state-of-the-art LLMs: GPT-3.5 and GPT-4 by OpenAI, Claude by Anthropic, Gemini by DeepMind, Mistral, and LLaMA 2 by Meta (OpenAI, 2024b,a; AI, 2024a; DeepMind, 2024; AI, 2023b, 2024b). These models span a range of architectures and optimization strategies, from lightweight systems such as Mistral (7–12 billion parameters) to large-scale models like GPT-4, which is estimated to exceed 1 trillion parameters.

All responses were generated using each model's official API with default settings; no parameters were modified to ensure consistent testing conditions and eliminate tuning-related bias.

For a detailed comparison of model specifications, including parameter counts, availability (free, paid, or both), release sources, and distinctive model features, see Table 2.

3.2 Rationality Component

We define rationality as the extent to which an agent's decisions conform to logical and probabilistic principles under uncertainty, following dual-process theories from cognitive science and decision theory (Stanovich, 2011; Kahneman, 2011). Unlike intelligence, which reflects raw cognitive ability, rationality reflects adherence to normative reasoning standards, particularly in abstract or unfamiliar contexts.

To operationalize this construct, we use four well-established behavioral tasks known to elicit systematic reasoning failures in human cognition: the Wason Selection Task (Wason, 1968) tests deductive logic and conditional reasoning; the Conjunction Fallacy Task (Tversky and Kahneman, 1983) measures probabilistic misjudgment; the Stereotype-Based Base Rate Neglect (Tversky and

Table 3: Examples of Rationality Tests with Descriptions and Examples

Test Type	Description	Example
Wason Selection Task	Evaluate deductive reasoning by identifying conditions that falsify a rule.	Rule: "If a card has a vowel on one side, it must have an even number on the other." Cards: A, K, 4, 7. Which cards to turn? (Correct: A and 7).
Conjunction Fallacy Task	Tests probabilistic reasoning, avoiding errors where conjunctions are judged more probable.	Linda is 31, outspoken, and concerned with social justice. Which is more likely: (1) Linda is a bank teller, or (2) Linda is a bank teller and a feminist? (Correct: (1)).
Stereotype Base Rate Neglect	Measures reliance on stereotypes over statistical information.	A group has 30 engineers and 70 lawyers. A randomly chosen person is quiet and systematic. Is this person an engineer or a lawyer? (Correct: Lawyer).
Diagnostic Base Rate Neglect	Assesses reliance on anecdotal cues over statistical reasoning in diagnostic scenarios.	A medical test is 99% accurate. 1% of the population has the disease. If you test positive, what's the probability you have the disease? (Low: $\sim$50%).

Table 4: Definitions of the legal reasoning tasks evaluated.

Task	Definition
Civil Procedure	The body of law governing the processes and rules courts follow in civil lawsuits, including how cases are filed, tried, and appealed.
Constitutional Law	The area of law that interprets and applies the U.S. Constitution, governing the relationships between the government and individuals and dividing powers among government branches.
Contracts	The branch of law that deals with agreements between parties, including creating, enforcing, and breaching legally binding agreements.
Criminal Law	The area of law defining criminal offenses and the legal process for prosecuting and defending against charges, including arrest, trial, and sentencing.
Evidence	The rules and principles that govern what information can be presented in court to prove or disprove facts in a legal proceeding.
Real Property	Land and anything permanently attached to it, such as buildings and the rights associated with land ownership.
Torts	The area of law dealing with civil wrongs or injuries, providing remedies for individuals harmed by the actions or omissions of others.

Kahneman, 1974) highlights conflicts between statistical and stereotypical rationale; and the Diagnostic Base Rate Neglect (Barbey and Sloman, 2007) evaluates Bayesian inference with diagnostic cues.

Together, these tasks capture key aspects of rationality, including logical validity, probabilistic coherence, and bias resistance. We aim to assess answer accuracy and the *stability* of reasoning across repeated queries. These tests serve as diagnostic tools for evaluating alignment with normative standards of reasoning, rather than as evidence of human-like cognition. For details about each test type, including its description and a representative example, see Table 3.

3.3 Legal Reasoning Component

We compiled multiple-choice questions from the Multistate Bar Examination (MBE) for the legal reasoning component, covering seven key domains: Civil Procedure, Constitutional Law, Contracts, Criminal Law and Procedure, Evidence, Real Property, and Torts. These questions were carefully selected to mirror the complexities that legal practitioners encounter in real-world scenarios, ensuring an assessment that goes beyond factual recall to include higher-order reasoning (of Bar Examiners, 2024; Healy, 2004; Hutchins, 1928; Hetland, 1965; Ackerman, 1989). For more details about the Multistate Bar Examination (MBE) and its sections, refer to Table 4.

3.4 Measuring Consistency and Stability

We employ several statistical measures to quantify the stability of model responses across multiple runs. The *Test-Retest Consistency Score (TRCS)* captures the proportion of identical re-

sponses across repeated trials, offering insight into how consistently a model behaves when presented with the same prompt. The *Intraclass Correlation Coefficient (ICC)* assesses the reliability of model outputs across multiple iterations (Weir, 2005).

In addition, we use *Analysis of Variance (ANOVA)* to determine whether differences in consistency scores across models are statistically significant (Kaufmann, 2010).

These metrics provide a robust framework for evaluating consistency in rationality and legal tasks. A high rate of *Answer Reversibility*—where the same model changes its response to the same prompt—suggests instability in reasoning and raises concerns about the model's reliability in professional and high-stakes applications.

4 Consistency Analysis: Measuring Inconsistency in Legal and Rationality Tasks

This section evaluates large language models' consistency and generalization capabilities (LLMs) regarding test-retest stability, statistical validation, and model scaling effects.

4.1 Measuring Test-Retest Stability and Intraclass Correlation

Test-Retest Consistency Score (TRCS) analysis reveals substantial response variability across repeated trials, with a standard deviation of 0.23, indicating that even advanced large language models (LLMs) lack stable reasoning structures. Intra-Class Correlation (ICC) scores further confirm this instability, with low mean ICC values across all models, demonstrating their failure to maintain consistent decision-making.

Despite achieving high accuracy in legal tasks, models still exhibit low ICC scores, suggesting that their performance is driven by statistical pattern exploitation rather than genuine legal reasoning. Similarly, rationality tasks such as the Wason Selection and Conjunction Fallacy Test show low ICC values, reinforcing the conclusion that LLMs lack internal coherence. These inconsistencies raise serious concerns about deploying LLMs in professional and legal domains, where stable and principled reasoning is essential. Unlike human experts, LLMs can produce erratic responses, leading to unpredictable decision-making that may carry significant legal and ethical consequences. Figure 1 presents a visualization of TRCS and ICC scores across different LLMs. The bar chart represents TRCS scores, measuring response consistency, while the line plot shows ICC scores, indicating the agreement between repeated responses.

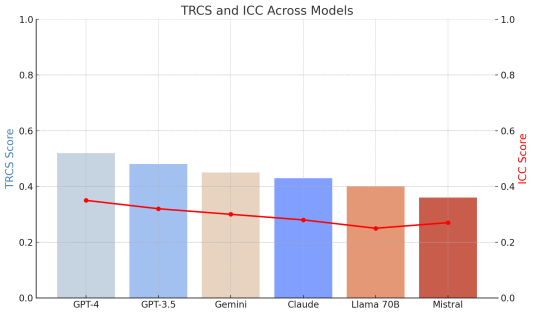

Figure 1: **TRCS and ICC Scores Across LLMs.** This figure presents each model's Test-Retest Consistency Score (TRCS) and Intraclass Correlation Coefficient (ICC). TRCS (blue bars) quantifies the proportion of identical responses across repeated trials, while ICC (red line) captures the consistency of model behavior across related prompts within each task domain.

4.2 Statistical Analysis of Consistency Differences

To examine response stability further, we conducted a one-way Analysis of Variance (ANOVA) and Levene's test across rationality and legal reasoning tasks. The ANOVA results ($F(3.65), p = 0.0059$) indicate statistically significant differences in consistency scores across models ($p < 0.05$), confirming that at least one model exhibits distinct response stability patterns. However, Levene's test ($F(1.38), p = 0.245$) suggests that response variability remains statistically comparable across models ($p > 0.05$).Despite GPT-4 demonstrating greater consistency than other models, inconsistencies persist. Figure 2 further illustrates the variation in consistency scores across task types and models, emphasizing the domain-specific performance of each model.

4.3 Model Scale and Consistency

Our findings reveal that Claude, GPT-4, and LLaMA 3.2 exhibit the highest consistency scores, outperforming smaller models like GPT-3.5 and Mistral. However, model size alone does not guarantee consistency. Despite being significantly larger than Claude, GPT-4 demonstrates more significant response variability. This suggests that alignment techniques and optimization strategies are more critical in stability than sheer scale.

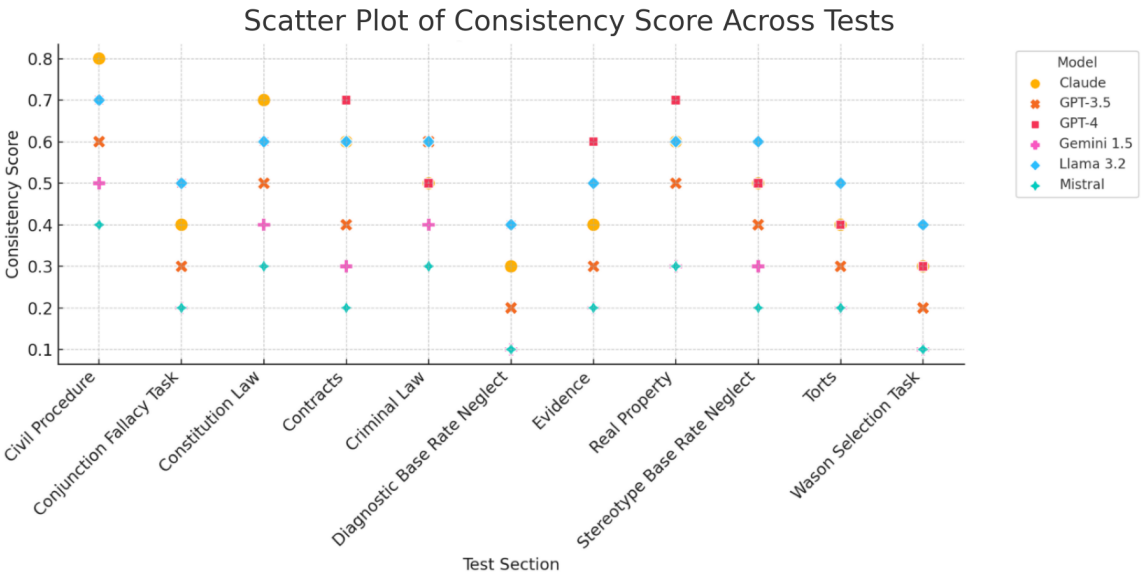

Figure 2: **Consistency Scores Across Test Sections.** This scatter plot compares the consistency scores of six large language models (LLMs)—Claude, GPT-4, GPT-3.5, Gemini 1.5, LLaMA 3.2, and Mistral—across various legal and rationality test sections. Each point represents a model's score on a specific task. The plot illustrates performance variability across domains and highlights that model size is not the sole determinant of consistency. Notably, Claude and LLaMA 3.2 demonstrate higher consistency across multiple tasks, while GPT-4 shows greater variability despite its scale.

Claude prioritizes consistency, whereas GPT-4's variability highlights the impact of different training methodologies. Similarly, LLaMA 3.2 achieves high consistency despite being smaller than GPT-4, reinforcing that structured learning approaches can enhance stability without extreme parameter scaling. Figure 3 compares consistency scores across models, illustrating performance variability and emphasizing that model size alone is not the primary determinant of stability.

Table 5 further demonstrates GPT-4's inconsistencies in legal reasoning, where identical prompts yielded different responses and distinct logical explanations.

4.4 Interpreting TRCS and ICC Divergences

TRCS and ICC capture different dimensions of consistency: TRCS measures how reliably a model repeats the same answer to identical prompts, while ICC assesses stability across related but distinct items within a domain.

Our findings show that models can exhibit high TRCS yet low ICC. For example, GPT-4 consistently answered specific legal prompts (TRCS = 1.00), but showed weak generalization across the full legal set (ICC = 0.42). In rationality tasks such as base rate neglect, GPT-4 was inconsistent both within items (TRCS = 0.50) and across functions (ICC = 0.31), revealing unstable probabilistic reasoning.

These results highlight that relying on TRCS alone may overstate model reliability. A combined analysis of TRCS and ICC offers a more accurate view of reasoning stability, essential for deploying LLMs in high-stakes domains like law and rationality.

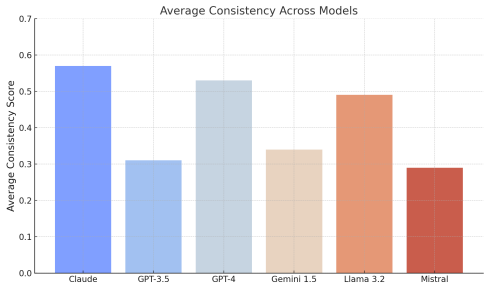

Figure 3: **Consistency Scores Across Models.** This figure compares average consistency scores across six language models. It illustrates performance variability and emphasizes that model size is not the key determinant of stability. Claude and GPT-4 show higher consistency, while smaller models like LLaMA 3.2 outperform larger ones like GPT-3.5, challenging assumptions about scale and reliability.

Table 5: GPT-4 Prompt, Answer Choices, Correct Answer, and Explanation

Prompt	Answer Choices	First Attempt	Second Attempt
A factory foreman was suspected of murdering the rival of a local union leader. After his arrest, the police asked his aunt if she knew where firearms could be found. She retrieved a pistol from the bedroom. Ballistics linked the pistol to the crime, and the foreman's fingerprints were found on it. The grand jury indicted the foreman after hearing this Evidence. He seeks to quash the indictment.	(A) No, because the Evidence was offered before a grand jury, not a court. (B) No, because the pistol was obtained by a private citizen, not the police. (C) Yes, because the police did not have probable cause to seize the pistol. (D) Yes, because the foreman's aunt was acting as an agent of the police when she obtained the pistol.	(B) No, because the pistol was obtained by a private citizen, not the police. *Explanation:* The exclusionary Rule only applies to government actors, and here, the foreman's aunt, a private citizen, voluntarily retrieved the pistol without coercion or direction from the police. Thus, the exclusionary Rule does not apply.	(D) Yes, because the foreman's aunt was acting as an agent of the police when she obtained the pistol. *Explanation:* This scenario involves the Fourth Amendment, and the question is whether the Evidence obtained by the supervisor's aunt, who acted as an agent of the police, should be excluded.
Correct Answer: (B) No, because the pistol was obtained by a private citizen, not the police (Barbri, 2024).			

4.5 Fluctuations and Inconsistencies in Diagnostic Reasoning

We observed marked fluctuations in model performance on the diagnostic base rate neglect task. LLMs frequently solved complex Bayesian inference problems, yet failed on simpler base rate items. In several instances, the selected answers directly contradicted the models' own explanations, revealing internal inconsistencies or stochastic behavior.

These findings echo the work of Macmillan-Scott and Musolesi (Macmillan-Scott and Musolesi, 2024a), who documented similar failures in GPT-4—successfully completing advanced mathematical tasks while erring on basic arithmetic and pattern recognition. Such discrepancies underscore a fundamental limitation: current LLMs lack stable internal reasoning mechanisms. This instability challenges the adequacy of accuracy as a standalone metric for evaluating model reliability.

4.6 The Importance of Consistency in High-Stakes Domains

Consistency is crucial for AI models, especially in legal and medical fields where reliability is essential. Contradictory responses undermine credibility, making AI unsuitable for decision-making that requires precision and coherence. In legal reasoning, inconsistencies in LLM-generated arguments can lead to unreliable analyses, while fluctuating diagnoses in medical applications pose significant risks. Moreover, inconsistent outputs complicate AI governance by reducing transparency and accountability. Ensuring AI models are high-performing, stable, fair, and interpretable is critical for their adoption in regulated environments (Stanovich, 2011; Flanagan and Alfonso, 2013).

4.7 Challenges in Cross-Domain Generalization

Unlike humans, LLMs struggle to generalize reasoning across domains and develop stable frameworks that transfer knowledge effectively (Gentner and Markman, 1997; Holyoak and Thagard, 2012). Legal professionals, for instance, demonstrate strong rationality beyond the law, applying structured reasoning to novel contexts (Weinreb, 2005; Gick and Holyoak, 1980). In contrast, LLMs exhibit response inconsistencies and logical drift in broader rationality tests despite excelling in domain-specific tasks. Their reliance on pattern recognition hinders cross-domain knowledge transfer (Marcus, 2020; Bender et al., 2021). This inconsistency raises ethical concerns, particularly in law and governance. LLMs often produce contradictory responses to similar prompts, posing risks in high-stakes applications (Weidinger et al., 2021; Hendrycks et al., 2021). Unlike human experts, they operate in isolated silos, leading to fragmented and unreliable decision-making.

4.8 Strategies for Enhancing LLM Generalization and Stability

Consistency is essential for deploying large language models (LLMs) in high-stakes legal reasoning and rational decision-making fields. While techniques like Reinforcement Learning with Human Feedback (RLHF) and Retrieval-Augmented Generation (RAG) have improved accuracy and alignment with human preferences, they are not explicitly designed to ensure stable responses across repeated prompts.

One approach to improving consistency is adapting RLHF to penalize contradictory outputs. For

instance, if a model provides different answers to the same legal question across multiple runs, its reward should be reduced. This can be implemented through a two-step process: first, generate several completions for the same prompt; second, select the best output with expert references or previous completions. This encourages more stable and reproducible behavior (Ouyang et al., 2022; Christiano et al., 2017).

Retrieval-based models can also be enhanced for consistency. Instead of relying on a single retrieved passage, the model can gather multiple related documents and generate responses that reflect agreement across these sources. For example, a prompt concerning the Fourth Amendment might retrieve precedents such as *Katz v. United States* and *Carpenter v. United States* to ensure alignment with established case law. This redundancy helps reduce hallucinations and mitigates drift in output across trials (Lewis et al., 2020; Izacard and Grave, 2021).

Another promising direction is a model-agnostic method called *Self-Consistency Regularization (SCR)*. Although not implemented in this study, SCR penalizes variability in model responses to the same prompt by applying entropy-based penalties or majority voting across Monte Carlo samples. Since SCR operates during inference, it can be integrated into existing systems without retraining, making it a scalable and complementary tool for enhancing stability.

Cognitively inspired strategies also show potential. Hierarchical reasoning allows models to decompose complex questions into smaller, interpretable components (Bengio et al., 2019; Lake et al., 2017). Meta-reasoning modules can assess the internal coherence of outputs before finalization, while lightweight memory systems enable models to recall prior responses, supporting consistency across interactions.

Holtzman et al. (Holtzman et al., 2021) demonstrate that even state-of-the-art models frequently produce inconsistent outputs when asked the same prompt repeatedly. This underscores the importance of treating consistency as a formal evaluation criterion. In our work, we adopt this view through the use of the *Test-Retest Consistency Score (TRCS)* (Lin et al., 2022; Mitchell, 2023), which measures how often a model returns the same answer across repeated trials. Broader adoption of such metrics is critical for developing reliable and trustworthy LLMs.

Taken together, these strategies—reward-modified RLHF, redundancy-aware retrieval, SCR, and cognitively inspired mechanisms—offer a promising path toward building models that are accurate but also stable, interpretable, and dependable in professional applications where consistency is essential.

5 Conclusion

In this study, we investigated fairness in large language models (LLMs) beyond accuracy, emphasizing consistency as a crucial factor for their reliability. We evaluated reasoning across six different LLMs and found that all exhibited a lack of consistency, posing a significant barrier to their deployment in high-stakes environments. Our analysis highlights the limitations of current models in maintaining stable reasoning patterns, which are essential for legal, medical, and other critical applications. Addressing these challenges requires further research into improving model consistency, incorporating adaptive learning strategies, and enhancing transparency to ensure trustworthy AI decision-making.

Limitations and Future Work

While this study highlights essential consistency challenges in large language models, several limitations remain. The analysis relies on a 30-day evaluation period, which, although offering a representative snapshot of consistency trends, may not fully capture the variability in model responses across different conditions or over time. To maintain feasibility, we adopted a streamlined protocol that ensured balanced representation across model tiers and tasks. However, this limited timeframe may not account for longer-term fluctuations or backend updates that could influence model performance.

Second, the evaluation focuses primarily on legal and rationality tasks. Although these are critical high-stakes areas—and rationality is foundational to legal reasoning—the exclusion of other domains limits the generalizability of our findings. In future work, we plan to expand the evaluation framework to include medical reasoning, computer science, and financial decision-making to assess cross-domain consistency better.

Third, we did not consider the potential effects of fine-tuning or retrieval-augmented generation (RAG) (Lewis et al., 2020; Izacard and Grave, 2021), which may enhance consistency and reli-

ability. Future research should explore these approaches and human-in-the-loop strategies to improve performance stability and trustworthiness in real-world applications.

Finally, a key direction for future work is to examine whether rationality generalizes across domains. While rationality is critical for legal reasoning, it remains unclear whether LLMs can consistently apply rational principles—such as logical validity, probabilistic reasoning, and resistance to cognitive biases—in unfamiliar contexts. Understanding the scope and limits of rational generalization is essential for developing more robust and trustworthy AI systems.

Acknowledgments

We are grateful to Roberto Porras, a licensed attorney in New York from the University of Virginia School of Law, and Moriah Bisewski, a licensed attorney in Texas from the University of Houston Law Center, for their valuable input and dialogue while preparing this work. Their insights helped us refine the legal components of this work and the precision of our analysis.

We also thank Omar Kamal for his assistance in revising the codebase and implementing the necessary modifications that supported our experimental framework.

Finally, we are thankful to Dr. Suha Beydoun for her guidance on the rationality tasks and her contributions to ensuring the fairness and quality of the evaluation process.

References

Bruce Ackerman. 1989. Constitutional politics/constitutional law. *Harvard Law Review*, 99(4):453–538.

Anthropic AI. 2024a. Claude 2 and constitutional ai: A safety-focused approach.

Meta AI. 2023a. Llama 3: Open large language models. Available at https://ai.meta.com/llama.

Meta AI. 2024b. Llama 3.2: Advancements in open-source ai.

Mistral AI. 2023b. Mistral 7b: A high-performance lightweight model.

Mistral AI. 2023c. Mistral 7b: Lightweight and efficient large language model. Available at https://mistral.ai.

Anthropic. 2023. Claude 2: Safety-centric large language model. Available at https://www.anthropic.com.

Alon Avrahami, Ron Shamir, and Daniel Cohen. 2023. Uncertainty in ai-based financial forecasting: The role of input sensitivity and response stability. *Journal of Finance and AI*, 9(1):55–72.

Aron K Barbey and Steven A Sloman. 2007. Base-rate respect: From ecological rationality to dual processes. *Behavioral and Brain Sciences*, 30(3):241–254.

Barbri. 2024. *Barbri Simulated MBE*. SIM MBE.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, et al. 2021. On the dangers of stochastic parrots: Can language models be too big? *FAccT*.

Yoshua Bengio et al. 2019. Meta-transfer learning for few-shot learning. *arXiv preprint arXiv:1910.10736*.

Maximilian Binz and Eric Schulz. 2023. Using cognitive science to evaluate ai rationality. *Proceedings of AAAI*.

Michael Bommarito and Daniel Martin Katz. 2022. Gpt-4 and the bar exam: What do we learn? *MIT Computational Law Report*.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Alexander P. Burgoyne, Michael J. Kane, and David Z. Hambrick. 2023. Understanding diagnostic base rate neglect: A meta-analytic review. *Journal of Experimental Psychology: General*, 152(1):112–131.

Ilias Chalkidis, Ion Androutsopoulos, and Nikolaos Aletras. 2021. Legal-bert: Pretrained transformers for legal text mining. *arXiv preprint arXiv:2103.11121*.

Jonathan H. Choi et al. 2023. Chatgpt goes to law school. *arXiv preprint arXiv:2304.00067*.

Paul F Christiano, Jan Leike, Tom Brown, et al. 2017. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems (NeurIPS)*.

A. Curcio. 2002. Bar exam performance and law school pedagogy: A critical examination. *Legal Education Review*, 15:143–167.

Arion Das, Asutosh Mishra, Amitesh Patel, Soumilya De, V Gurucharan, and Kripabandhu Ghosh. 2024. Can llms faithfully generate their layperson-understandable'self'?: A case study in high-stakes domains. *arXiv preprint arXiv:2412.07781*.

DeepMind. 2023. Gemini 1: Multimodal large language model. Available at https://deepmind.com/gemini.

Google DeepMind. 2024. Gemini 1.5 pro: High-performance language model.

Emilio Ferrara, Stefano Cresci, and David Lazer. 2023. Hallucination and response drift in large language models: Implications for fairness and bias. *Nature Machine Intelligence*, 5(8):678–692.

Dawn P. Flanagan and Vincent C. Alfonso. 2013. *The Cattell-Horn-Carroll Theory and Evidence-Based Assessment*. Springer.

Lin Gao, Mei Zhang, and Xinyu Li. 2023. Evaluating consistency in ai-assisted medical diagnostics: A multi-trial study. *Journal of Medical AI Research*, 12:202–215.

Dedre Gentner and Arthur B Markman. 1997. Structure-mapping in analogy and similarity. *American psychologist*, 52(1):45.

Mary L Gick and Keith J Holyoak. 1980. Analogical problem solving. *Cognitive psychology*, 12(3):306–355.

Carol Goforth. 2015. Why the bar examination fails to raise the bar. *Ohio NUL Rev.*, 42:47.

Thomas Healy. 2004. Criminal law: A systematic approach. *American Criminal Law Review*, 41(1):135–156.

B. Heidemann. 2020. Strategies for mastering the mbe: Legal analysis and test-taking techniques. *Legal Education Quarterly*, 33:200–225.

Dan Hendrycks, Collin Burns, Saurav Kadavath, et al. 2021. Aligning ai with shared human values. *arXiv preprint arXiv:2008.02275*.

Leif M. Hetland. 1965. The definition of real property. *Minnesota Law Review*, 49:731–748.

Ari Holtzman, Peter West, Chandra Bhagavatula, and Yejin Choi. 2021. Surface form competition: Why the highest probability answer isn't always right. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 769–784.

Keith J Holyoak and Paul Thagard. 2012. *Analogy and relational reasoning*. Oxford University Press.

Robert M. Hutchins. 1928. Some observations on evidence and evidence teaching. *Columbia Law Review*, 28(3):432–444.

Gautier Izacard and Edouard Grave. 2021. Leveraging passage retrieval with generative models for open domain question answering. *arXiv preprint arXiv:2101.00294*.

Thanmay Jayakumar, Fauzan Farooqui, and Luqman Farooqui. 2023. Large language models are legal but they are not: Making the case for a powerful legalllm. *arXiv preprint arXiv:2311.08890*.

P. Johnson and C. White. 2021. Bar exam and legal competence: Evaluating the testing framework. *Law Ethics Journal*, 28:189–210.

Daniel Kahneman. 2011. *Thinking, Fast and Slow*. Farrar, Straus and Giroux, New York.

Mina Katsumi and Brian S. Liu. 2023. Assessing ai's legal knowledge: Beyond standardized testing. *AI Law Review*.

Stefan Kaufmann. 2010. Analysis of variance (anova). In *Wiley Interdisciplinary Reviews: Computational Statistics*. John Wiley & Sons.

Brenden M Lake, Tomer D Ullman, Joshua B Tenenbaum, and Samuel J Gershman. 2017. Building machines that learn and think like people. *Behavioral and Brain Sciences*, 40:e253.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems (NeurIPS)*.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Truthfulqa: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958*.

Michael A. Livermore and Ashley Southall. 2023. Law in the age of ai: Evaluating gpt-4's legal performance. *Harvard Journal of Law and Technology*.

Oliver Macmillan-Scott and Mirco Musolesi. 2024a. (ir)rationality and cognitive biases in large language models. *Royal Society Open Science*, 11(6):240255.

Olivia Macmillan-Scott and Mirco Musolesi. 2024b. (ir)rationality and cognitive biases in large language models. *Royal Society Open Science*, 11(6):240255.

Gary Marcus. 2020. The next decade in ai: Four steps towards robust artificial intelligence. *arXiv preprint arXiv:2002.06177*.

Eric Martínez. 2024. Re-evaluating gpt-4's bar exam performance. *Artificial Intelligence and Law*, pages 1–24.

Melanie Mitchell. 2023. Abstraction and generalization in ai and human cognition. *Artificial Intelligence Review*, 56:251–272.

J. Nystrom. 2013. Bar exam structure and its impact on legal education. *Journal of Legal Studies*, 29:312–330.

National Conference of Bar Examiners. 2024. Mbe subject matter outline. Available: https://www.ncbex.org/exams/mbe/.

OpenAI. 2023a. Chatgpt-3.5: Reliable language model. Available at https://openai.com.

OpenAI. 2023b. Gpt-4 and bar exam results. Accessed: YYYY-MM-DD.

OpenAI. 2023c. Gpt-4 performance benchmarks. Accessed: YYYY-MM-DD.

OpenAI. 2023d. Gpt-4 technical report. https://openai.com/research/gpt-4.

OpenAI. 2023e. Gpt-4 technical report. Accessed: YYYY-MM-DD.

OpenAI. 2024a. Gpt-3.5 turbo: Reliable and versatile ai model.

OpenAI. 2024b. Gpt-4o: Advancing ai reasoning and problem-solving.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, et al. 2022. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*.

Anika Rawal, Rishabh Gupta, and Sungjin Lee. 2023. Assessing logical stability and bias in ai decision-making systems. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(4):2850–2862.

Denise Riebe. 2006. A bar review for law schools: Getting students on board to pass their bar exams. *BrandeiS lJ*, 45:269.

Shibani Santurkar et al. 2023. Whose reasoning? evaluating llms on logical, causal, and probabilistic tasks. *Proceedings of NeurIPS*.

Patrick Schramowski, Dominik Stammer, and Volker Tresp. 2022. Large language models are not fair evaluators: An analysis of gpt-3's moral and ethical reasoning capabilities. *arXiv preprint arXiv:2203.10259*.

N. Schwartz. 2007. Success strategies for bar exam candidates. *Journal of Bar Studies*, 20:56–75.

M. Simkovic and F. McIntyre. 2015. Should the bar exam be redesigned? a statistical perspective. *Law Society Review*, 41:75–98.

Keith E. Stanovich. 2011. *Rationality and the Reflective Mind*. Oxford University Press, New York, NY.

Amos Tversky and Daniel Kahneman. 1974. Judgment under uncertainty: Heuristics and biases. *Science*, 185(4157):1124–1131.

Amos Tversky and Daniel Kahneman. 1983. Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, 90(4):293–315.

Meng Wang, Xiangrui Lin, et al. 2023. Towards reasoning in large language models: A survey. *arXiv preprint arXiv:2301.05698*.

Peter C. Wason. 1968. Reasoning about a rule. *Quarterly Journal of Experimental Psychology*, 20(3):273–281.

Laura Weidinger, Iason Gabriel, Amelia Glaese, Rishi Bommasani, et al. 2021. Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*.

Laura Weidinger et al. 2022. Ethical and social risks of large language models. *Advances in Neural Information Processing Systems*, 35:35412–35423.

Lloyd L Weinreb. 2005. Legal reason: The use of analogy in legal argument. *Cambridge University Press*.

Joseph P. Weir. 2005. Quantifying test-retest reliability using the intraclass correlation coefficient and the sem. *Journal of Strength and Conditioning Research*, 19(1):231–240.

Hao Zhong, Jieyu Tang, Tianyang Xu, et al. 2020. Does nlp benefit legal system? a case study of document representation. *arXiv preprint arXiv:2005.01647*.

Haoyang Zhong, Yujia Wang, and Wei Yang. 2023. Can large language models be consistent? evaluating stability in legal and policy decision-making. *AI Society*, 38:1123–1140.