

What Drives an LLM’s Legal Reasoning? Measuring Faithfulness at the Concept Level

Yassine Thlija^{1,2} Lena Held² Ivan Habernal²

Trustworthy Human Language Technologies

¹ Department of Computer Science, Technical University of Darmstadt

² Research Center Trustworthy Data Science and Security of the University Alliance Ruhr,
Faculty of Computer Science, Ruhr University Bochum

www.trusthlt.org

Abstract

Faithfulness assesses whether a model’s explanation reflects what actually drove its answer, which is different from assessing answer correctness. But does faithfulness also apply in expert domains such as law, where the relevant factors that decide the answer, also called concepts, are diffuse and only implicitly expressed? We measure faithfulness in two expert-annotated legal reasoning multiple-choice question answering (QA) datasets and show that existing methods are not tailor-made for the legal domain. We therefore introduce a detailed, per-concept measure which reveals how each concept is being relied on and expressed by a model. We test our method with a small, open-weight LLM and find that faithfulness is generally low on legal datasets and that simple supervised fine-tuning lowers faithfulness even further, despite giving plausible legal explanations. The code is publicly available.¹

1 Introduction

Explanations play an important role in understanding and following legal reasoning, as legal conclusions derive their legitimacy from justification through facts, doctrines, procedural rules, and competing interpretations. As a result, models are expected to produce explanations or step-by-step analyses alongside their final answers. However, a convincing legal explanation does not necessarily reveal the factors that actually drove the prediction.

In expert domains, like the legal domain, reading a convincing explanation for an answer can be very misleading when it is not true to the model’s actual internal reasoning. The explanation may contain domain-specific statements that are difficult to audit without legal knowledge.

The distinction is commonly described as the difference between plausibility and faithfulness (Jacovi and Goldberg, 2020). Plausibility concerns

whether an explanation appears convincing to a human reader, while faithfulness concerns whether the explanation accurately reflects the factors that influenced the model’s prediction.

In legal NLP, however, this distinction has largely gone unexamined. The primary focus has been task performance, with benchmarks (Guha et al., 2023; Chalkidis et al., 2022) and correctness metrics dominating the evaluation. Comparatively less attention has been paid to whether generated legal explanations faithfully reflect the model’s underlying decision process. Furthermore, it remains unclear whether adapting LLMs to produce legal explanations improves explanation faithfulness or merely improves explanation plausibility.

Both questions demand a way to measure faithfulness in a domain where concepts are diffuse and expressed implicitly. In this work, we investigate how faithfulness can be measured in the legal domain and which adaptations of existing methods are necessary. Building on the faithfulness framework of (Matton et al., 2025), we propose a faithfulness measure for legal question answering that compares a model’s answer prediction with the decisive factors emphasized in the explanation, accounting for diffuse and implicitly expressed concepts of legal reasoning. Additionally, we explore possibilities to obtain a fine-grained diagnostic which lets us gain a better understanding of the mismatch between what drives a model’s answer and its explanation. Lastly, we ask how faithfulness is influenced by supervised fine-tuning and investigate this by comparing the performance of a base model to a model fine-tuned for a legal reasoning task, to observe whether accuracy and faithfulness move together. We apply our method to two datasets, the Legal Argument Reasoning Task in Civil Procedure (Bongard et al., 2022) and the Multi-level Legal Reasoning dataset (Islam et al., 2025), in which each answer is paired with an explanation written by a legal expert.

¹<https://github.com/trusthlt/what-drives-legal-faithfulness>

2 Related work

2.1 Legal reasoning datasets

Several benchmarks have been proposed to evaluate legal reasoning capabilities. CaseHOLD (Zheng et al., 2021) frames legal reasoning as a multiple-choice task requiring models to identify the correct legal holding from a set of candidate options. LegalBench (Guha et al., 2023) broadens evaluation across diverse legal reasoning tasks. However, these benchmarks primarily focus on prediction accuracy and generally do not provide explicit reasoning steps explaining how answers are derived.

Some legal datasets do contain explicit reasoning. One such dataset is the Legal Argument Reasoning Task in Civil Procedure (Bongard et al., 2022), derived from *The Glannon Guide to Civil Procedure* (Glannon, 2019). Each instance contains a legal question together with a detailed analysis explaining why the correct answer follows from the relevant legal principles and why alternative answers are incorrect. More recently, the Multi-level Legal Reasoning dataset (Islam et al., 2025) introduced human-written and model-generated explanations at multiple levels of abstraction.

These explanation-rich datasets are particularly relevant for studying faithfulness because they provide explicit legal reasoning traces rather than only ground-truth answer labels.

2.2 Faithfulness and plausibility of explanations

Recent work has shown that natural language explanations produced by LLMs can appear convincing while failing to reflect the factors that actually drive model predictions. Turpin et al. (2023) demonstrate that chain-of-thought explanations can systematically omit causally relevant factors influencing model outputs. Similarly, Madsen et al. (2024) find substantial variation in the faithfulness of self-explanations across models, tasks, and explanation formats. Atanasova et al. (2023) further propose dedicated tests for evaluating the faithfulness of natural language explanations.

Research on explainability has also produced rationale-based approaches (Lei et al., 2016), benchmark-based evaluation frameworks such as ERASER (DeYoung et al., 2020), and perturbation or counterfactual methods for estimating causal influence (Chan et al., 2022; Liu et al., 2022). Surveys by Jacovi and Goldberg (2020) and Lyu et al.

(2024) provide broader overviews of the challenges involved in defining and measuring explanation faithfulness.

Most closely related to our work is the *Walk the Talk* framework (Matton et al., 2025), which compares concept importance inferred from model behavior with concept importance implied by generated explanations. While *Walk the Talk* has been evaluated on general reasoning tasks, its applicability to legal reasoning has not yet been studied. Recent work by McMillan et al. (2025) further used *Walk the Talk* to evaluate the faithfulness of medical reasoning models trained with different supervision and reinforcement-learning strategies. Their findings suggest that improvements in reasoning performance do not necessarily correspond to improvements in explanation faithfulness, motivating further investigation in expert domains such as law.

Although substantial progress has been made in both explanation faithfulness and legal reasoning, their intersection remains largely unexplored.

3 Methodology

We build on the *Walk the Talk* framework of Matton et al. (2025), which measures whether the factors that influence a model’s answer are also reflected in its explanation. We first introduce the terminology used by the framework and describe its original formulation. We then discuss why directly applying it to legal reasoning is difficult and introduce our adaptations for measuring and diagnosing faithfulness in legal question answering.

3.1 Terminology

Walk the Talk evaluates faithfulness in terms of *concepts*. A concept c_i represents a factual property of a question that may influence the model’s answer and that can take alternative values. For a question q , we denote its concepts by

$$\mathcal{C}(q) = \{c_1, \dots, c_n\}. \quad (1)$$

Each concept defines a set of possible *counterfactual values*. A counterfactual intervention changes the value of one concept while keeping the remaining question unchanged. For a concept c_i , we denote the resulting counterfactual questions by

$$\tilde{q}_{i,1}, \dots, \tilde{q}_{i,k_i}. \quad (2)$$

Counterfactuals allow us to observe whether changing a particular aspect of a question changes the

model’s answer. For example, consider the following simplified question:

A 17-year-old enters into a contract to purchase a car. Is the contract enforceable?

One concept is the *age of the contracting party*, whose counterfactual values might include 18 or 25. Another could be the *type of transaction*, with alternative values corresponding to legally meaningful variations of the original transaction. Replacing “17-year-old” with “25-year-old” creates a counterfactual question in which only the selected concept has been changed. If this intervention substantially changes the model’s answer, the model demonstrates behavioral reliance on that concept.

3.2 Walk the Talk

An *explanation* is the natural-language justification generated by the model for its answer. It represents what the model explicitly claims to be relevant to its decision. Faithfulness can therefore be studied by comparing two views of the same concept: how much changing it affects the model’s answer and how important the model presents it to be in its explanation.

Matton et al. (2025) formalize these two views as the *Causal Concept Effect* (CE) and the *Explanation-Implied Effect* (EE). CE captures a model’s behaviorally demonstrated reliance on a concept, while EE captures the reliance implied by its generated explanation.

Causal Concept Effect. For each concept c_i , Matton et al. (2025) compare the model’s answer distribution for the original question with the distributions obtained after counterfactual interventions. Matton et al.’s causal concept effect $\text{CE}^M(c_i)$ of concept c_i is the average Kullback–Leibler divergence $\mathbb{D}$ between these distributions:

$$\text{CE}^M(c_i) = \frac{1}{k_i} \sum_{j=1}^{k_i} \mathbb{D}(P(\cdot \mid \tilde{q}_{i,j}) \parallel P(\cdot \mid q)), \quad (3)$$

where k_i is the number of counterfactual questions for concept c_i . A larger CE indicates that interventions on the concept produce larger changes in the model’s answer distribution and therefore that the model relies more strongly on it.

Explanation-Implied Effect. The second quantity measures whether the same concept is reflected in the model’s explanation. Matton et al. (2025) treat this as a binary effect: a concept either appears

as a causal factor in the explanation or it does not. The resulting explanation-implied effect $\text{EE}^M(c_i)$, defined as

$$\text{EE}^M(c_i) = \frac{1}{k_i} \sum_{j=1}^{k_i} P(c_i \in E \mid \bar{q}_{i,j}), \quad (4)$$

therefore captures whether the model’s stated reasoning acknowledges the concept that is being probed.

Causal Concept Faithfulness. For each question, CE and EE form two vectors over its concepts, denoted CE_q^M and EE_q^M , respectively. Matton et al. (2025) standardize both vectors within the question and estimate their relationship using a hierarchical regression. A faithful explanation should assign relatively high explanatory importance to concepts with relatively high causal effects. Their measure $\mathcal{F}^M(q)$ therefore captures whether demonstrated and stated concept importance move together by calculating the Pearson correlation ρ between CE^M and EE^M

$$\mathcal{F}^M(q) = \rho(\text{CE}_q^M, \text{EE}_q^M). \quad (5)$$

3.3 Challenges for legal reasoning

The intuition behind Walk the Talk transfers naturally to legal reasoning: if a fact substantially influences a model’s legal conclusion, we would expect its explanation to reflect that influence. Its original formulation, however, poses several problems when applied directly to legal questions.

First, legally relevant concepts are often diffuse. They are not necessarily isolated entities or explicit phrases, but factual relationships whose legal importance emerges from how several parts of a scenario interact. A useful concept must therefore capture an editable factual dimension of the legal scenario, rather than merely identifying whether a particular term occurs.

Second, legal explanations rarely divide concepts cleanly into “used” and “not used.” A fact may be mentioned only as background, used as supporting evidence, or form the decisive reason for reaching a conclusion. Treating all of these cases as the same binary explanatory effect discards precisely the differences in importance that characterize legal argumentation.

Finally, the original faithfulness measure is primarily concerned with the existence of the linear covariance of CE and EE across concepts. It therefore cannot distinguish an explanation that tracks

the model’s reliance proportionally from one that flattens or exaggerates the differences between concepts, as long as the ordering and linear fit are comparable. For our purposes, this distinction is itself informative: not only we want to identify whether the same concepts are relatively important, but also whether the explanation accurately represents *how much* the model relies on each of them.

3.4 Our faithfulness measure

We retain the central comparison between behavioral and explicit reliance, but adapt how concepts, CE, EE, and their agreement are measured. Figure 1 summarizes the resulting pipeline.

Concepts and Counterfactuals. We operationalize a concept as an editable factual dimension of a legal scenario that may influence the answer. Concepts are extracted using only the introductory context and question text (see a full example from the dataset in Appendix B). We withhold answer candidates and expert analyses so that concept extraction reflects the factual structure of the scenario rather than information derived from the correct answer.

For every concept, we generate alternative values and construct counterfactual questions by modifying only that concept. Concepts and counterfactuals must be grounded in the original scenario and permit an isolated intervention while preserving the remaining facts. In this way, each concept defines a counterfactual value space through which its influence on the model’s prediction can be probed. An example of counterfactual replacements can be found in Appendix C.

Causal Concept Effect. We estimate the answer distribution $P(\cdot | q)$ directly from the likelihoods assigned by the target model to the complete answer candidates. We normalize these likelihoods by answer length and calibrate them against a content-free prompt to reduce preferences arising from answer wording and prompt structure (Zhao et al., 2021). The calibrated scores are converted into a distribution over the answer candidates using a temperature-scaled softmax. The complete procedure is described in Appendix D.5.

Instead of KL divergence, we measure the effect of an intervention using total variation (TV) distance:

$$\text{CE}_{\text{raw}}(c_i) = \frac{1}{k_i} \sum_{j=1}^{k_i} \text{TV}(P(\cdot | q), P(\cdot | \tilde{q}_{i,j})), \quad (6)$$

where

$$\text{TV}(P, Q) = \frac{1}{2} \sum_a |P(a) - Q(a)|. \quad (7)$$

TV is symmetric and bounded in $[0, 1]$, giving the effect a direct interpretation as the amount of probability mass shifted by an intervention. Because the largest possible shift depends on the original answer distribution, we normalize by the maximum attainable TV distance:

$$\text{CE}(c_i) = \frac{\text{CE}_{\text{raw}}(c_i)}{1 - \min_a P(a | q)}. \quad (8)$$

CE therefore expresses how much of the available shift in the model’s answer distribution is produced by intervening on the concept.

Explanation-Implied Effect. To capture the different roles concepts can play in legal explanations, we replace binary concept presence with graded importance. An external judge model assigns each concept one of five scores:

- 0 = not present / irrelevant
- 1 = mentioned only as background
- 2 = relevant supporting factor
- 3 = major reason
- 4 = decisive or contrastive reason.

We evaluate the concept in the explanation of the original question and in the explanations generated for its corresponding counterfactuals. Let $\mathcal{E}_i$ denote these explanations and $J(c_i, e) \in \{0, \dots, 4\}$ the judge’s rating. We define

$$\text{EE}(c_i) = \frac{1}{4|\mathcal{E}_i|} \sum_{e \in \mathcal{E}_i} J(c_i, e). \quad (9)$$

EE is therefore bounded in $[0, 1]$ and captures *how strongly* a concept is represented in the model’s explicit reasoning, rather than only whether it occurs.

Faithfulness. Following Matton et al. (2025), we measure faithfulness by the association between a model’s behavioral reliance on concepts and the importance assigned to those concepts in its explanations. For a question q , we define the concept-level effect vectors

$$\mathbf{CE}_q = (\text{CE}(c_1), \dots, \text{CE}(c_n)) \quad (10)$$

and

$$\mathbf{EE}_q = (\text{EE}(c_1), \dots, \text{EE}(c_n)). \quad (11)$$

We compute the Pearson correlation coefficient (PCC) between these vectors:

$$\mathcal{F}(q) = \rho(\mathbf{CE}_q, \mathbf{EE}_q). \quad (12)$$

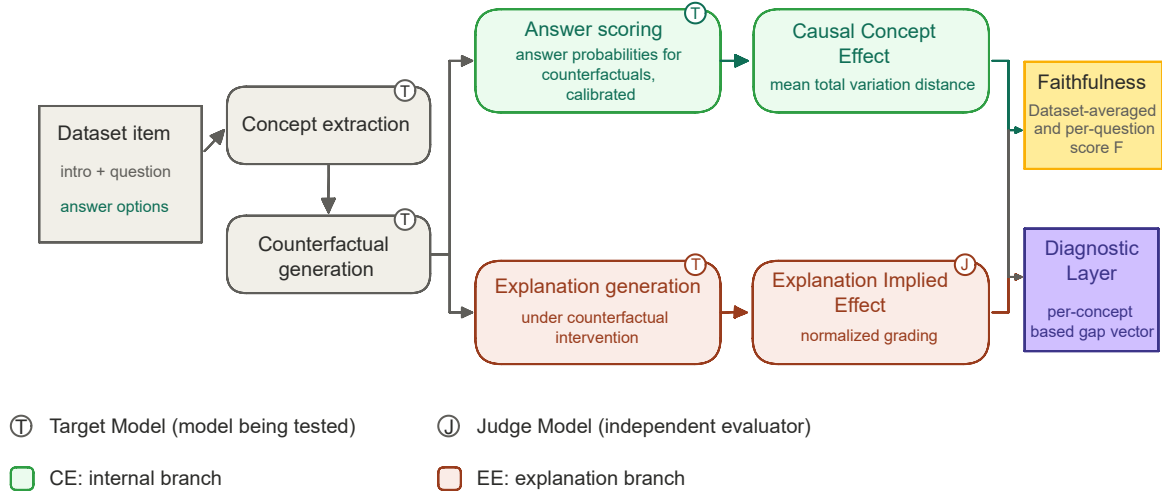


Figure 1: Our faithfulness evaluation pipeline. Concepts are extracted from the question and introductory context, and counterfactual values are generated for each concept. Behavioral reliance is estimated from changes in the target model’s (T) answer distribution under these interventions (CE). Explicit reliance is estimated by an external judge model (J) from the importance assigned to each concept in generated explanations (EE). Faithfulness measures the agreement between the two quantities.

A high positive correlation indicates that concepts on which the model demonstrates greater behavioral reliance also tend to receive greater importance in its explanation. Conversely, a low or negative correlation indicates a mismatch between the concepts that influence the model’s answer and those emphasized in its explicit reasoning.

3.5 Concept-level diagnostics

A question-level score tells us whether the two views of the model’s reasoning agree overall, but it does not tell us *where* they disagree. This distinction is particularly important for our goal of diagnosing legal reasoning rather than reducing faithfulness to a single score.

For each concept, we therefore compute the signed difference vector

$$\text{gap}(c_i) = \text{EE}(c_i) - \text{CE}(c_i). \quad (13)$$

A positive gap means that the explanation assigns more importance to a concept than is supported by the model’s behavior under intervention. We refer to this as *over-emphasis*. A negative gap means that the model’s answer demonstrably relies on a concept more strongly than its explanation acknowledges, which we refer to as *under-emphasis*. A gap close to zero indicates local agreement.

The question-level faithfulness score and concept-level gaps therefore answer different questions. The former measures whether the explanation agrees with the model’s demonstrated reason-

ing overall; the latter identifies which concepts are responsible for a mismatch and in which direction. This lets us move from asking whether a legal explanation is faithful to examining *how* and *where* its account of the model’s reasoning fails.

4 Experimental setup

4.1 Legal reasoning datasets

The first dataset is the *Legal Argument Reasoning Task in Civil Procedure* (Bongard et al., 2022), constructed from *The Glannon Guide to Civil Procedure* (Glannon, 2019). The dataset contains multiple-choice legal questions paired with expert-written analyses explaining why the correct answer follows from the relevant legal principles while also discussing why the remaining answer options are incorrect. Following the original split provided by Bongard et al. (2022), we use the training split for fine-tuning and merge the development and test splits into a single held-out evaluation set to increase the number of evaluation examples.

As a second benchmark, we use the *Multi-level Legal Reasoning* dataset (Islam et al., 2025). The dataset contains 150 multiple-choice questions covering diverse areas of international law, including human rights, humanitarian law, environmental law, and international trade. Each instance includes a human-written legal explanation together with model-generated reasoning traces from multiple perspectives (general, judge, and advocate). Be-

cause our framework evaluates reasoning rather than factual legal knowledge, we first annotate each question using GPT-5.4-mini as either a *reasoning* question or a *knowledge* question. Questions classified as knowledge questions were excluded, as they typically contain few meaningful reasoning concepts for our concept-based faithfulness metric. This filtering yields a subset of 100 reasoning questions. Since the original dataset is relatively small, we combine the provided training and test splits into a single evaluation set.

Together, these datasets allow us to evaluate our framework across different legal domains and styles of legal reasoning.

4.2 Target model

The model to be tested for faithfulness is called the target model. We use Magistral Small 1.2, released as mistralai/Magistral-Small-2509 (Mistral-AI et al., 2025), throughout our experiments, because it is an open-weight model designed for reasoning tasks. Its open weights support reproducible experimentation and permit parameter-efficient fine-tuning, while its reasoning capabilities make it suitable for studying the relationship between generated legal explanations and model behaviour.

We evaluate the base model on both legal reasoning datasets by generating concepts, counterfactuals and extracting answer probabilities. We report answer accuracy, defined as the proportion of questions for which the model’s selected multiple-choice answer matches the gold answer, alongside the explanation faithfulness metric introduced in Section 3.4. We additionally evaluate a fine-tuned version of Magistral Small on the held-out test split of the Glannon dataset. This model is referred to as the *Glannon Reasoner* throughout the paper.

For evaluation, the model is prompted to generate a legal explanation and to end its response with a single answer letter in the format Final answer: X. The full inference prompts and generation parameters are provided in Appendix D.2. Figure 2 shows extracted concepts and counterfactuals for an example question.

4.3 Judge model

Following Matton et al. (2025), who use a frontier OpenAI model as a judge for estimating EE, we use GPT-5.4-mini to evaluate the explanations produced by the target model. For each explanation, the judge receives the explanation and the set of ex-

Question (paraphrased). Eban brings suit against Lorenzo for interference with business relations. Eban’s lawyer, Darrow, calls Lorenzo’s lawyer, Sadecki, and tells her that he is filing suit that afternoon. In which of the following scenarios may the case proceed without formal service of process?

Candidates.

(A) Darrow files the complaint with the district court clerk with an affidavit swearing that he informed Sadecki of the case and mailed her a copy of the complaint. (B) Darrow files the complaint and mails a copy of it by first class mail to Lorenzo with two copies of a proper request for waiver. He files an affidavit with the court attesting to those acts. **(C) Darrow files the complaint and mails a copy of it by certified mail to Lorenzo with two copies of a proper request for waiver. He does not send a summons, signed and sealed by the court, with the waiver request. He receives the signed waiver form back and files it with the court.** (D) Darrow files the complaint and mails a copy of it by certified mail to Lorenzo with two copies of a proper request for waiver. He receives the green postal receipt back in the mail. Darrow files the postal receipt with the court

Extracted concepts & counterfactuals (Magistral-Small)

C1: Claim			
breach of contract	defamation	fraud	negligence
C2: Communication			
discuss the lawsuit filing	negotiate the lawsuit filing		
argue about the lawsuit filing			
C3: Defendant			
Sadecki	Darrow		
C4: Plaintiff			
Darrow	Lorenzo	Sadecki	

Figure 2: An example question taken from the Glannon dataset with concepts and counterfactuals extracted by Magistral-Small.

tracted legal concepts and assigns each concept an importance score on the five-point scale described in Subsection 3.4.

The judge model is used only to grade the target model’s explanations. It does not extract concepts, generate counterfactual inputs, or predict answers. We query the model with temperature set to zero. The complete judge prompt and API configuration are provided in Appendix D.1.

4.4 Fine-tuning for legal explanations

To investigate whether fine-tuning on expert-written legal analyses improves explanation faithfulness, we fine-tune Magistral Small on the Glannon training split. Each training instance contains the introductory legal text, question, candidate answers, and expert-written analysis. The model is trained to generate the expert-style legal analysis followed by the predicted answer in the format

Final answer: X.

We fine-tune using Low-Rank Adaptation (LoRA) on 161 training examples and evaluate the resulting *Glannon Reasoner* on a disjoint held-out set of 48 examples. Complete training details, hyperparameters, and preliminary fine-tuning configurations are reported in Appendix D.3.

5 Results

5.1 Model performance vs. model faithfulness

Table 1 reports answer accuracy together with our question-level faithfulness score ($\mathcal{F}$) and the original faithfulness score of Matton et al. (2025) ($\mathcal{F}^M$).

Across both datasets, answer accuracy and explanation faithfulness show substantially different patterns. Magistral-Small achieves its highest accuracy on Multilevel (87.0%), while its faithfulness remains considerably lower ($\mathcal{F} = 0.455$). On Glannon, accuracy is similar between the training and held-out test sets (63.0% and 62.5%, respectively), as are their faithfulness scores ($\mathcal{F} = 0.325$ and 0.355). These results suggest that stronger answer performance does not necessarily correspond to greater agreement between the concepts influencing a model’s predictions and those reflected in its explanations.

Model	Acc.	$\mathcal{F}$	$\mathcal{F}^M$
<i>Glannon (train)</i>			
Magistral-Small	.630	.325	.224
<i>Glannon (test)</i>			
Magistral-Small	.625	.355	.281
Glannon Reasoner	.604	.289	.177
<i>Multilevel</i>			
Magistral-Small	.870	.455	.270

Table 1: Answer accuracy, mean question-level faithfulness $\mathcal{F}$, and the original score $\mathcal{F}^M$ of Matton et al. (2025), per dataset.

Fine-tuning on expert-written Glannon explanations does not improve either answer accuracy or explanation faithfulness on the held-out test set. The Glannon Reasoner’s accuracy decreases from 62.5% to 60.4%, while $\mathcal{F}$ decreases from 0.355 to 0.289 and $\mathcal{F}^M$ from 0.281 to 0.177. Thus, in this setting, training the model to imitate expert-written legal explanations does not improve the alignment between the concepts emphasized in its explanations and those that influence its predictions.

Concept	CE	EE	gap	Verdict
Claim	.008	.050	.042	<i>faithful</i>
Communication	.093	.500	.407	<i>over-emphasis</i>
Defendant	.037	.750	.713	<i>over-emphasis</i>
Plaintiff	.021	.125	.104	<i>faithful</i>
$\mathcal{F} = .519$				$\mathcal{F}^M = 0.358$

Table 2: Matching per-concept scores to the example question provided in Figure 2. A concept is tagged with *over-emphasis* or *under-emphasis* if $|\text{gap}| > 0.15$, otherwise it is considered *faithful*.

5.2 Concept-level diagnostics

Aggregate faithfulness scores do not reveal which concepts are responsible for a mismatch between a model’s behavior and its explanation. Table 2 therefore illustrates our concept-level diagnostics for one representative question from the held-out Glannon test set.

The model extracts four concepts and generates several counterfactual values for each. Intervening on these concepts yields their causal concept effects (CE), while the judge model estimates their importance in the generated explanations to obtain the explanation-implied effects (EE). The signed gap, $\text{EE} - \text{CE}$, then indicates whether a concept is emphasized in the explanation proportionally to its demonstrated influence on the model’s prediction.

The example reveals substantial differences across concepts. In particular, *Defendant* receives a high explanation-implied effect ($\text{EE} = 0.750$) despite a much smaller causal effect ($\text{CE} = 0.037$), resulting in a gap of 0.713. Similarly, *Communication* has an EE of 0.500 but a CE of only 0.093, yielding a gap of 0.407. Both concepts are therefore classified as *over-emphasized*. In contrast, *Claim* and *Plaintiff* show considerably smaller gaps (0.042 and 0.104, respectively) and are classified as *faithful* under our threshold of $|\text{gap}| \leq 0.15$.

At the question level, the example obtains a faithfulness score of $\mathcal{F} = 0.519$ and a score of $\mathcal{F}^M = 0.358$ under the original measure of Matton et al. (2025). While these aggregate scores summarize the overall correspondence between behavioral and explicit reliance, the concept-level gaps localize the source and direction of the mismatch. In this example, they show that the disagreement is primarily driven by the model assigning substantially greater explanatory importance to *Defendant* and *Communication* than their counterfactual effects suggest.

6 Discussion

Accuracy is not faithfulness. Across both datasets, answer accuracy and explanation faithfulness appear largely decoupled. Magistral-Small achieves 87% accuracy on the Multi-level dataset despite obtaining a very low faithfulness score, while fine-tuning on expert-written legal explanations reduces faithfulness under both our metric and the correlation-based measure of [Matton et al. \(2025\)](#). These findings reinforce that correct answers and convincing explanations do not necessarily reflect the concepts that actually drive a model’s prediction.

Supervised fine-tuning learns explanation style.

Fine-tuning on expert legal analyses produces explanations that more closely resemble expert writing without improving faithfulness. This suggests that supervised fine-tuning primarily learns the distribution and structure of expert explanations rather than aligning them with the model’s underlying decision process. Explanation quality and explanation faithfulness should therefore be viewed as complementary objectives rather than interchangeable ones.

Concept-level diagnostics expose failure modes.

Dataset-level faithfulness scores summarize overall behavior but do not reveal why a model is unfaithful. Our concept-level analysis introduced in Section 3.4 identifies which concepts are *over-emphasized* or *under-emphasized* relative to their demonstrated influence on the prediction. Figure 2 and the complete per-question analyses in Appendix A illustrate how similar overall faithfulness scores can arise from very different reasoning failures.

Implications for legal reasoning evaluation.

Legal benchmarks typically emphasize prediction accuracy, yet our results suggest that answer correctness alone provides an incomplete picture of model behaviour. Reporting explanation faithfulness alongside accuracy offers a more informative evaluation of legal reasoning systems and raises the question of whether faithfulness itself should become an optimization objective, rather than merely an evaluation criterion.

7 Conclusion

We adapted a concept-level faithfulness score to the legal domain and showed that the way faith-

fulness is measured matters as much as the score itself. A single dataset-level number cannot distinguish between different failure modes and even though [Matton et al. \(2025\)](#)’s framework is well-suited to its original setting, legal reasoning with its diffuse and often correlated concepts demands a more distinguished view. Our adapted measure scores each concept directly and reports the difference between demonstrated and claimed reliance, separating concepts that are relied on but under-expressed from those that are overexpressed yet relied on. To account for the more nuanced way that concepts are used in legal explanations, we score on a graded, categorical scale rather than a binary causal presence.

Faithfulness on both legal reasoning datasets was generally low. Simple supervised fine-tuning on expert analyses makes the model produce plausible legal-sounding explanations, yet faithfulness slightly dropped, showing that the plausible and faithful explanations are not necessarily tied together. And while this is not a surprising outcome of using supervised fine-tuning, it highlights the need to look beyond accuracy alone when evaluating language models for legal reasoning. We hope our per-concept diagnostic serves as a step towards treating faithfulness as an additional evaluation target in the legal domain.

Because our measure can rate individual explanations, it could open up the possibility of using faithfulness directly as a training signal, rather than only as an evaluation metric. We evaluated a single model on two datasets, and for broader claims about legal reasoning faithfulness in LLMs, more models and datasets that genuinely require legal reasoning capabilities would need to be put under scrutiny. Whether faithfulness can be optimized as a training target, and whether it can improve a model’s legal reasoning remains an open challenge worth pursuing.

Limitations

Our measure captures faithfulness, which can be interpreted as self-consistency between demonstrated and claimed reliance, rather than correctness. A model can be faithful while giving a wrong answer. The frame work also applies only to multiple-choice legal datasets, since the answer distribution needs to be extracted over fixed answer options. Its reliability depends on the quality of the extracted concepts and counterfactuals and a weaker model

may land lower faithfulness scores simply for failing to produce meaningful concepts. Extracted concepts do not have to be legally sound, since the measure concerns consistency between claimed and demonstrated reliance rather than legal correctness. Furthermore, the measure depends on the quality of the judge model. Scoring the explanations of every counterfactual variant is also costly, as it requires an additional judge call for each one. [Matton et al. \(2025\)](#) avoid this by scoring only the original explanation and approximating the distribution with an estimator. Finally, our fine-tuning results are based on a small split of the Glannon dataset ([Bongard et al., 2022](#)) (48 questions), which is not the same split as the original dataset, since the original version was formatted as a natural language inference task with a question, a single answer candidate and a binary label. Thus the scores cannot be compared directly.

Acknowledgments

This work has been partly supported by the Research Center Trustworthy Data Science and Security (<https://rc-trust.ai>), one of the Research Alliance centers within the <https://uaruhr.de>, and by the German Research Foundation as part of the ECALP project (HA 8018/2-1).

References

- Pepa Atanasova, Oana-Maria Camburu, Christina Lioma, Thomas Lukasiewicz, Jakob Grue Simonsen, and Isabelle Augenstein. 2023. [Faithfulness tests for natural language explanations](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 283–294, Toronto, Canada. Association for Computational Linguistics.
- Leonard Bongard, Lena Held, and Ivan Habernal. 2022. [The legal argument reasoning task in civil procedure](#). In *Proceedings of the Natural Legal Language Processing Workshop 2022*, pages 194–207, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.
- Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael Bommarito, Ion Androutsopoulos, Daniel Katz, and Nikolaos Aletras. 2022. [LexGLUE: A benchmark dataset for legal language understanding in English](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4310–4330, Dublin, Ireland. Association for Computational Linguistics.
- Chun Sik Chan, Huanqi Kong, and Liang Guanqing. 2022. [A comparative study of faithfulness metrics for model interpretability methods](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5029–5038, Dublin, Ireland. Association for Computational Linguistics.
- Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. 2020. [ERASER: A benchmark to evaluate rationalized NLP models](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4443–4458, Online. Association for Computational Linguistics.
- Joseph W. Glannon. 2019. *The Glannon Guide to Civil Procedure: Learning Civil Procedure Through Multiple-Choice Questions and Analysis*, 4 edition. Wolters Kluwer, New York.
- Neel Guha, Julian Nyarko, Daniel Ho, Christopher Ré, Adam Chilton, Aditya K, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel Rockmore, Diego Zambrano, Dmitry Talisman, Enam Hoque, Faiz Surani, Frank Fagan, Galit Sarfaty, Gregory Dickinson, Haggai Porat, Jason Hegland, and 21 others. 2023. [Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 44123–44279. Curran Associates, Inc.
- Kang He, Yinghan Long, and Kaushik Roy. 2024. [Prompt-based bias calibration for better zero/few-shot learning of language models](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 12673–12691, Miami, Florida, USA. Association for Computational Linguistics.
- Mst Rafia Islam, Azmine Toughshik wasi, and Umong Sain. 2025. [Legal reasoning dataset with multilevel human and model-annotated explanations](#).
- Alon Jacovi and Yoav Goldberg. 2020. [Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness?](#) In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4198–4205, Online. Association for Computational Linguistics.
- Zhengbao Jiang, Jun Araki, Haibo Ding, and Graham Neubig. 2021. [How can we know when language models know? on the calibration of language models for question answering](#). *Transactions of the Association for Computational Linguistics*, 9:962–977.
- Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2016. [Rationalizing neural predictions](#). In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 107–117, Austin, Texas. Association for Computational Linguistics.
- Yibing Liu, Haoliang Li, Yangyang Guo, Chenqi Kong, Jing Li, and Shiqi Wang. 2022. Rethinking attention-model explainability through faithfulness violation test. In *Proceedings of the 39th International Conference on Machine Learning*, pages 13807–13824. PMLR.

- Qing Lyu, Marianna Apidianaki, and Chris Callison-Burch. 2024. [Towards faithful model explanation in NLP: A survey](#). *Computational Linguistics*, 50(2):657–723.
- Andreas Madsen, Sarath Chandar, and Siva Reddy. 2024. [Are self-explanations from large language models faithful?](#) In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 295–337, Bangkok, Thailand. Association for Computational Linguistics.
- Katie Matton, Robert Ness, John Gutttag, and Emre Kiciman. 2025. [Walk the talk? measuring the faithfulness of large language model explanations](#). In *International Conference on Learning Representations*, volume 2025, pages 73212–73277.
- Teague McMillan, Gabriele Dominici, Martin Gjoreski, and Marc Langheinrich. 2025. [Towards transparent reasoning: What drives faithfulness in large language models?](#) In *NeurIPS 2025 Workshop on Evaluating the Evolving LLM Lifecycle: Benchmarks, Emergent Abilities, and Scaling*.
- Mistral-AI, :, Abhinav Rastogi, Albert Q. Jiang, Andy Lo, Gabrielle Berrada, Guillaume Lample, Jason Rute, Joep Barmentlo, Karmesh Yadav, Kartik Khandelwal, Khyathi Raghavi Chandu, Léonard Blier, Lucile Saulnier, Matthieu Dinot, Maxime Darrin, Neha Gupta, Roman Soletskyi, Sagar Vaze, and 82 others. 2025. [Magistral](#). *Preprint*, arXiv:2506.10910.
- Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. 2023. Language models don’t always say what they think: unfaithful explanations in chain-of-thought prompting. In *Proceedings of the 37th International Conference on Neural Information Processing Systems, NIPS ’23*, Red Hook, NY, USA. Curran Associates Inc.
- Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. [Calibrate before use: Improving few-shot performance of language models](#). In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 12697–12706. PMLR.
- Lucia Zheng, Neel Guha, Brandon R. Anderson, Peter Henderson, and Daniel E. Ho. 2021. [When does pre-training help? assessing self-supervised learning for law and the casehold dataset of 53,000+ legal holdings](#). In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law, ICAIL ’21*, page 159–168, New York, NY, USA. Association for Computing Machinery.

A Detailed results

Question	Concept	CE	EE	gap	$\mathcal{F}$	$\mathcal{F}^M$
31	court type	0.012	0.438	0.426 ↑	0.593	0.244
	dispute type	0.070	1.000	0.930 ↑		
	employee status	0.007	0.375	0.368 ↑		
	immunity claim	0.006	1.000	0.994 ↑		
	legal claim	0.012	0.600	0.588 ↑		
	organization	0.001	0.000	−0.001		
	location					
33	fishing location	0.005	1.000	0.995 ↑	0.455	0.230
	fishing vessel	0.001	0.050	0.049		
	origin					
	protestor	0.003	0.062	0.060		
	pursuit location	0.037	0.917	0.880 ↑		
	seizing authority	0.002	0.250	0.248		
	vessel status	0.249	0.875	0.626 ↑		
35	cost	0.034	0.750	0.716 ↑	0.390	0.375
	country	0.000	0.000	−0.000		
	label design	0.132	0.875	0.743 ↑		
	product	0.000	0.100	0.100		
	purpose	0.000	0.850	0.850 ↑		
	requirement	0.000	0.950	0.950 ↑		
41	claim	0.010	0.300	0.290 ↑	0.519	−0.414
	country	0.000	0.400	0.400 ↑		
	domestic dairy	0.000	0.750	0.750 ↑		
	products					
	imported dairy	0.113	0.750	0.637 ↑		
52	products				0.732	0.575
	civilian presence	0.273	1.000	0.727 ↑		
	country	0.090	0.150	0.060		
	situation	0.215	0.875	0.660 ↑		
	treaty status	0.413	0.812	0.400 ↑		
	weapon	0.331	1.000	0.669 ↑		

Table 3: Detailed faithfulness results for Magistral-Small on selected instances of the Multi-level legal reasoning test set.

Question	Concept	CE	EE	gap	$\mathcal{F}$	$\mathcal{F}^M$
11. Investment in litigation	Carmichael claims	0.039	0.550	0.511 ↑	0.450	0.328
	Carmichael claims basis	0.091	0.812	0.721 ↑		
	Faithful defenses	0.049	0.333	0.285 ↑		
	Faithful response	0.039	0.250	0.211		
	claimed rate of return	0.129	0.375	0.246		
	investment amount	0.012	0.000	−0.012		
	judge finding 2	0.080	0.917	0.837 ↑		
	legal standard	0.055	1.000	0.945 ↑		
	trial type	0.010	0.312	0.302 ↑		
9. Removing the objections	court type	0.052	0.833	0.781 ↑	0.494	0.205
	defendant	0.100	0.875	0.775 ↑		
	jurisdiction	0.309	0.875	0.566 ↑		
	approach					
	objections raised	0.029	0.812	0.784 ↑		
	plaintiff	0.042	0.000	−0.042		
	removal to	0.050	0.438	0.388 ↑		
	response to	0.453	0.875	0.422 ↑		
	complaint					
8. Technical fouls	time before answer due	0.018	0.350	0.332 ↑		
	claim	0.008	0.050	0.042	0.561	0.358
	communication	0.093	0.500	0.407 ↑		
	defendant	0.037	0.750	0.713 ↑		
	plaintiff	0.021	0.125	0.104		
9. Sufficient information	AE’s knowledge	0.522	0.938	0.416 ↑	0.566	0.311
	Adair’s employment status	0.022	0.250	0.228		
	accident involved	0.015	0.000	−0.015		
	allegation	0.045	1.000	0.955 ↑		
7. Comedy of errors	Emporia Supreme Court decision	0.112	1.000	0.888 ↑	0.643	0.120
	appeal status	0.043	0.625	0.582 ↑		
	court	0.072	0.583	0.511 ↑		
	injured parties	0.016	0.062	0.046		
	timing of appeal	0.063	0.917	0.854 ↑		
	hearing					
	trial judge’s decision	0.085	0.917	0.832 ↑		
	type of action	0.084	0.250	0.166		

Table 4: Detailed faithfulness results for Magistral-Small on selected instances of the Glannon test set.

Question	Concept	CE	EE	gap	$\mathcal{F}$	$\mathcal{F}^M$
7. If at first you don't succeed	Ervin's next step	0.017	0.500	0.483 ↑	0.056	-0.111
	Ito's defense in Acadia	0.019	1.000	0.981 ↑		
	Ito's response to first court ruling	0.047	0.900	0.853 ↑		
	claim	0.011	0.050	0.039		
	claimant	0.071	0.083	0.013		
	defendant	0.039	0.750	0.711 ↑		
	first court	0.035	0.625	0.590 ↑		
	first court objection	0.054	0.950	0.896 ↑		
	first court ruling	0.043	1.000	0.957 ↑		
10. Second thoughts	defendant	0.014	0.125	0.111	0.340	0.405
	defendant's status	0.010	0.000	-0.010		
	federal statute	0.003	0.562	0.559 ↑		
	plaintiff	0.030	0.150	0.120		
	plaintiff's basis for claim	0.007	1.000	0.993 ↑		
	plaintiff's claim	0.008	0.375	0.367 ↑		
	plaintiff's motion	0.057	1.000	0.943 ↑		
	plaintiff's status	0.017	0.438	0.420 ↑		
	time elapsed	0.031	0.500	0.469 ↑		
10. Gag order	basis	0.120	0.875	0.755 ↑	0.561	0.442
	conduct	0.032	1.000	0.968 ↑		
	role	0.008	0.125	0.117		
7. A venue medley	accident location	0.032	1.000	0.968 ↑	0.647	0.168
	defendant	0.007	0.833	0.826 ↑		
	defendant's employer	0.004	0.562	0.558 ↑		
	defendant's job	0.006	0.000	-0.006		
	defendant's state	0.009	0.500	0.491 ↑		
	employer's principal place of business	0.008	0.650	0.642 ↑		
	filing location	0.072	1.000	0.928 ↑		
	plaintiff	0.003	0.350	0.347 ↑		

Table 5: Detailed faithfulness results for the Glannon Reasoner on selected instances of the Glannon test set.

B Example Legal Reasoning Instance

To illustrate the structure of the legal reasoning data used in our experiments, we provide one complete example from the Legal Argument Reasoning Task in Civil Procedure (Bongard et al., 2022). Each instance consists of an introductory legal text, a question, multiple answer candidates, the correct answer, and an expert-written analysis explaining the answer.

Question Title. *1. Moving Marla*

Introductory Text. Under federal diversity jurisdiction a citizen of one state may sue a citizen of another in federal court, even though her claim arises under state law, if she has a colorable claim for more than \$75,000. The state citizenship of a person—as opposed to that of a corporation—is determined by her domicile, that is, the most recent state where she has (1) resided with (2) the intent to remain indefinitely. The “residence” requirement is easily satisfied. Staying overnight in a hotel or a tent will establish “residence” in a state. It is the intent-to-remain-indefinitely prong that gives students problems. A person intends to remain indefinitely in a state if she is residing in the state on an open-ended basis, without the intent to leave at a definite time or on the occurrence of a definite event. You don’t have to swear allegiance forever to a state to acquire domicile there; you only need to reside there “indefinitely,” that is, on an open-ended basis. If a party is living in a state without definite plans to leave, the domicile test regards that state as her “home.” She is there, not as a visitor, but as a citizen. She is, psychologically speaking, at home there, rather than passing through. She may choose to move on, as we all may, but at the moment she has no plans to do so. Remember that, until the two prongs coincide in a new state, your old domicile continues, whether you plan to return to that state or not. If Acari, from Hawaii, leaves for a one-year job acting in a play in California, planning to go to New York afterwards, he remains domiciled in Hawaii, even if he swears that he will never return to Hawaii. He hasn’t acquired a domicile in California, because he doesn’t plan to stay there indefinitely. He hasn’t acquired one in New York either, since he doesn’t reside there yet. Domicile doctrine abhors a vacuum, so it holds that Acari keeps his Hawaii domicile until the two prerequisites come together in another state. In analyzing the question below, assume that the court applies

the reside-with-intent-to-remain-indefinitely test, and consider where Marla has established a residence with the requisite “indefinite” intent.

Question. Marla, who grew up in Montana, moved to Colorado after high school to enter a two-year program for hair stylists at the Denver Beauty School. She wasn’t sure if she really wanted to be a stylist, but she was anxious to get away from home, and her parents agreed to foot the bill, so off she went. She figured she’d stay if she liked it, and get a job as a stylist afterwards, in Denver or elsewhere in the West (including Montana). Or she would leave the program if she didn’t like it and look for work, hopefully in Denver. She took an apartment on a six-month lease. After moving to Denver, Marla

Possible Answers.

- A. remains domiciled in Montana, because the program is only for two years.
- B. remains domiciled in Montana, because she may return there to work as a stylist.
- C. remains domiciled in Montana, because she was domiciled there before she left for Denver.
- D. is domiciled in Colorado, because she resides there with the intent to remain indefinitely.

Correct Answer. D. Marla is domiciled in Colorado because she resides there with the intent to remain indefinitely.

C Counterfactual Examples

Examples for counterfactuals produced by Magistral-small are shown in Figure 3 for the same questions as Figure 2.

D Hyperparameters and Experiment Details

D.1 Judge Model Details

We use GPT-5.4-mini as the judge model for estimating the Explanation-Implied Effect (EE). The judge is accessed through the OpenAI Responses API and is queried using greedy decoding with temperature set to zero.

For each generated explanation, the judge receives:

- the explanation generated by the target model,
- the extracted legal concepts,

Original question Eban brings suit against Lorenzo for interference with business relations. Eban’s lawyer, Darrow, calls Lorenzo’s lawyer, Sadecki, and tells her that he is filing suit that afternoon. In which of the following scenarios may the case proceed without formal service of process?

Generated counterfactuals (Magistral-Small)

C1: Claim

breach of contract: Eban brings suit against Lorenzo for **breach of contract**. Eban’s lawyer, Darrow, calls Lorenzo’s lawyer, Sadecki, and tells her that he is filing suit that afternoon. In which of the following scenarios may the case proceed without formal service of process?

defamation: Eban brings suit against Lorenzo for **defamation**. Eban’s lawyer, Darrow, calls Lorenzo’s lawyer, Sadecki, and tells her that he is filing suit that afternoon. In which of the following scenarios may the case proceed without formal service of process?

fraud: Eban brings suit against Lorenzo for **fraud**. Eban’s lawyer, Darrow, calls Lorenzo’s lawyer, Sadecki, and tells her that he is filing suit that afternoon. In which of the following scenarios may the case proceed without formal service of process?

negligence: Eban brings suit against Lorenzo for **negligence**. Eban’s lawyer, Darrow, calls Lorenzo’s lawyer, Sadecki, and tells her that he is filing suit that afternoon. In which of the following scenarios may the case proceed without formal service of process?

C2: Communication

Darrow and Sadecki discuss the lawsuit filing: Eban brings suit against Lorenzo for interference with business relations. **Darrow and Sadecki discuss the lawsuit filing**. In which of the following scenarios may the case proceed without formal service of process?

Darrow and Sadecki negotiate the lawsuit filing: Eban brings suit against Lorenzo for interference with business relations. **Darrow and Sadecki negotiate the lawsuit filing**. In which of the following scenarios may the case proceed without formal service of process?

Darrow and Sadecki argue about the lawsuit filing: Eban brings suit against Lorenzo for interference with business relations. **Darrow and Sadecki argue about the lawsuit filing**. In which of the following scenarios may the case proceed without formal service of process?

C3: Defendant

Sadecki: Eban brings suit against **Sadecki** for interference with business relations. Eban’s lawyer, Darrow, calls Lorenzo’s lawyer, Sadecki, and tells her that he is filing suit that afternoon. In which of the following scenarios may the case proceed without formal service of process?

Darrow: Eban brings suit against **Darrow** for interference with business relations. Eban’s lawyer, Darrow, calls Lorenzo’s lawyer, Sadecki, and tells her that he is filing suit that afternoon. In which of the following scenarios may the case proceed without formal service of process?

C4: Plaintiff

Darrow: **Darrow** brings suit against Lorenzo for interference with business relations. **Darrow’s** lawyer, Darrow, calls Lorenzo’s lawyer, Sadecki, and tells her that he is filing suit that afternoon. In which of the following scenarios may the case proceed without formal service of process?

Lorenzo: **Lorenzo** brings suit against Lorenzo for interference with business relations. **Lorenzo’s** lawyer, Darrow, calls Lorenzo’s lawyer, Sadecki, and tells her that he is filing suit that afternoon. In which of the following scenarios may the case proceed without formal service of process?

Sadecki: **Sadecki** brings suit against Lorenzo for interference with business relations. **Sadecki’s** lawyer, Darrow, calls Lorenzo’s lawyer, Sadecki, and tells her that he is filing suit that afternoon. In which of the following scenarios may the case proceed without formal service of process?

Figure 3: The example question from Figure 2 with the full set of counterfactuals generated by Magistral-Small. Each counterfactual substitutes a single concept value and holds the remainder of the question fixed.

- instructions describing the five-point importance scale defined in Subsection 3.4.

The judge independently assigns an importance score to each extracted concept. These scores are then used to compute the Explanation-Implied Effect (EE). The complete judge prompt is included in the supplementary material accompanying our code release.

The judge model is used exclusively for explanation evaluation and is never used for concept extraction, counterfactual generation, answer prediction, or fine-tuning.

D.2 Target Model Details

All experiments use Magistral Small 1.2 (mistralai/Magistral-Small-2509) (Mistral-AI et al., 2025).

During inference, the model receives:

- the introductory legal passage,
- the legal question,
- all answer candidates.

The model is instructed to generate a complete legal explanation before terminating with

Final answer: X

where X denotes the selected answer option. Generation uses:

- maximum context length: 4096 tokens,
- maximum generated tokens: 2048,
- greedy decoding ($T = 0$).

The complete inference prompts are shown below.

System prompt

You are an expert in US Civil Procedure.

Given an introductory legal text, a question, and multiple possible answers, produce a complete legal explanation that

- identifies the legally correct answer,
- explains why the correct answer is correct,
- discusses the relevant incorrect options,
- avoids repetition.

The final line must be

Final answer: X

User prompt

Introductory text:

{intro}

Question:

{question}

Possible answers:

{answers}

Write the legal explanation and end with the required final answer line.

D.3 Fine-Tuning Details

We fine-tune Magistral Small using parameter-efficient Low-Rank Adaptation (LoRA). The model is trained on 161 examples from the Glannon training split and evaluated on a disjoint held-out set of 48 examples.

Training Objective. Each training instance contains an introductory legal text, a multiple-choice question, its candidate answers, and the corresponding expert-written analysis. The model is trained to generate a complete legal explanation that identifies the correct answer, explains why it is correct,

and discusses relevant incorrect alternatives. Each response ends with the predicted answer in the format Final answer: X .

LoRA Configuration. LoRA is applied to the attention projections (q_{proj} , k_{proj} , v_{proj} , and o_{proj}) and the feed-forward projections ($gate_{\text{proj}}$, up_{proj} , and $down_{\text{proj}}$). The final Glannon Reasoner uses a LoRA rank of 64, a scaling parameter of 128, and a dropout rate of 0.05.

Hyperparameter	Value
Base model	Magistral Small 1.2
Training examples	161
Held-out test examples	48
Epochs	4
Learning rate	1×10^{-5}
LoRA rank (r)	64
LoRA scaling (α)	128
LoRA dropout	0.05
Maximum sequence length	4096
Maximum generated tokens	2048
Precision	bfloat16
Optimizer	AdamW (fused)
Learning-rate schedule	Linear
Weight decay	0
Warmup	0
Per-device batch size	1
Gradient accumulation	8
Effective batch size	8
Random seed	42
Hardware	NVIDIA H200 (140 GB)
Training time	2.93 hours

Table 6: Training configuration of the final Glannon Reasoner.

Reasoning Format. We initially experimented with placing the expert-written explanation inside [THINK] and [/THINK] delimiters, following Magistral’s reasoning format. This configuration resulted in lower held-out accuracy than training the model to generate the legal explanation directly without special reasoning delimiters. We hypothesize that, given the limited training data, fine-tuning content within the model’s reasoning delimiters may interfere with its pretrained reasoning format. We therefore use the no-[THINK] configuration for the final Glannon Reasoner. We treat this observation as an empirical result rather than evidence about the model’s latent reasoning process.

D.4 Development Experiments

We explored three fine-tuning configurations during model development.

The first configuration attempted to preserve

Magistral’s original reasoning format by supervising the model to generate its explanation inside [THINK] and [/THINK] delimiters.

The remaining experiments instead trained the model to generate legal explanations directly, without reasoning delimiters. This configuration consistently achieved better held-out performance and was therefore selected for the final Glannon Reasoner.

Table 7 reports the configurations explored during development. The reasoning-delimited experiment fine-tuned the model to produce its explanation within the pretrained [THINK] reasoning format. While this improved upon the initial model, training the model to emit the legal explanation directly without reasoning delimiters consistently achieved higher held-out accuracy. Consequently, all experiments reported in the main paper use the final Glannon Reasoner configuration.

D.5 Answer-Distribution Estimation and Debiasing

We estimate the model’s distribution over answer options from the sequence likelihoods of the complete candidate answers. This avoids relying on the probability of a first generated answer-label token, which may reflect preferences for particular labels rather than the semantic content of the candidates.

Let x denote the prompt containing the introductory passage, question, and answer candidates, and let y_i denote the full textual content of answer option i . We compute its base score as the length-normalized log-likelihood

$$u_i = \frac{1}{|y_i|} \log P(y_i | x), \quad (14)$$

where $|y_i|$ is the token length of the candidate. Length normalization prevents longer answers from receiving lower scores solely because their likelihood is accumulated over more tokens.

Null-prompt calibration. Following the calibration approach of He et al. (2024), we use a content-free prompt to reduce biases in the estimated answer scores. Sequence likelihoods may reflect intrinsic preferences for particular answer formulations or artifacts of the prompt structure independently of the question content. We estimate this bias using a content-free version of the prompt, x_{null} , in which the substantive question is replaced with N/A while the answer candidates and prompt structure are preserved. The corresponding null

score is

$$b_i = \frac{1}{|y_i|} \log P(y_i | x_{\text{null}}). \quad (15)$$

We obtain a calibrated score by subtracting a scaled version of this content-free preference:

$$s_i = u_i - \lambda b_i. \quad (16)$$

We set $\lambda = 0.04$, as the value that best preserved the original answer distributions in our calibration runs while still introducing a correction for content-free candidate preferences. Larger values gave the null-prompt correction greater influence and increasingly altered the original distributions.

Temperature scaling and probability normalization. After calibrating the candidate scores, we apply temperature scaling before the softmax, following the calibration procedure of Jiang et al. (2021). The calibrated scores are converted into a distribution over the answer candidates as

$$P(a_i | q) = \frac{\exp(s_i/\tau)}{\sum_{\ell=1}^m \exp(s_\ell/\tau)}, \quad (17)$$

where m is the number of answer candidates and $\tau > 0$ is the temperature.

The temperature controls the sharpness of the resulting distribution without changing the ordering of the candidate scores. As τ approaches zero, probability mass becomes increasingly concentrated on the candidate with the highest score, approaching a point mass on that candidate. Conversely, as τ increases, differences between candidate scores are compressed, and the distribution approaches a uniform distribution as $\tau \rightarrow \infty$. We use $\tau = 3$ to obtain a smoother distribution over candidate answers and use this value consistently across all questions.

The same scoring, calibration, and normalization procedure is applied independently to the original question and to every counterfactual question $\tilde{q}_{i,j}$. The resulting probabilities should be interpreted as relative preference distributions rather than calibrated Bayesian posteriors. They provide an observable estimate of the model’s preference over the answer candidates while controlling for answer length and content-free candidate preferences. The Causal Concept Effect in Subsection 3.4 is computed from shifts between these original and counterfactual distributions.

Configuration	Output	r	α	Drop.	Epochs	LR	Train. (%)	Runtime	Train	Test
Reasoning-delimited	[THINK]	32	64	0.00	3	1×10^{-5}	0.838	55.1 min	51.6	60.4
Direct explanation	No [THINK]	16	32	0.00	1	2×10^{-5}	0.421	49.8 min	52.2	54.2
Glannon Reasoner	No [THINK]	64	128	0.05	4	1×10^{-5}	1.662	176.1 min	64.0	70.8

Table 7: Summary of all fine-tuning experiments. The final Glannon Reasoner corresponds to the third configuration.