

On the Role of Supporting Evidence in LLM Reasoning: A Causal Analysis of Legal QA with Activation Patching

Dalmo Buzato Fernanda Sathler Caio Caldeira Isabela Santos
Evandro Cunha Pedro Calais Wagner Meira Jr. Anisio Lacerda

UFMG, Brazil

{dalmobuzato, caio.caldeira, isabelaribeiro, pcalais, meira, anisio}@dcc.ufmg.br
{fernandasathler, cunhae}@ufmg.br

Abstract

As large language models (LLMs) are increasingly considered for high-stakes legal applications, a central question for trust and auditability is whether their outputs reflect genuine reasoning over the supplied legal materials, or merely surface statistical patterns absorbed during pre-training. We address this question at the mechanistic level by applying activation patching, a causal-mediation technique, to four open-source multilingual LLMs on two bar-exam datasets: U.S. MBE and Brazilian OAB. For each multiple-choice question we contrast a gold supporting passage with an irrelevant-but-plausible counterfactual. We then surgically replace residual-stream activations across layers and token regions and measure recovery of the logit difference. Gold passages exert a measurable but model- and linguistic-dependent causal effect; integration concentrates in the upper layers (a late-binding signature); and the size of the effect is strongly mediated by the linguistic and discourse structure of the passage itself, not just its relevance. We introduce a four-cluster taxonomy that operationalises these regimes and identifies cases where retrieved evidence is causally decisive, inert, or actively harmful.

1 Introduction

Large language models (LLMs) are increasingly deployed in high-stakes domains such as medicine and law, where reasoning failures carry direct human cost (Kant et al., 2025; Griot et al., 2025). In legal practice in particular, the central question is no longer whether LLMs produce fluent text, but whether they reason over the legal materials they are given or simply pattern-match against text memorized during pretraining (Dahl et al., 2024). Distinguishing reasoning from retrieval is critical for AI safety, explainability, and the responsible deployment of retrieval-augmented systems across

jurisdictions where statute and jurisprudence vary widely (Nguyen et al., 2025).

We approach this question through *activation patching*, a mechanistic interpretability technique that implements a causal intervention over internal representations (Zhang and Nanda, 2023; Heimersheim and Nanda, 2024). Grounded in Judea Pearl’s interventionist causal inference (Pearl, 2001, 2012), it provides a natural fit for the question of whether retrieved evidence causally drives a legal answer; we develop the formal framing in Sec. 2.

We apply activation patching to four open-source multilingual LLMs, namely (i) Gemma-2-9B-it, (ii) Mistral-7B-v0.1, (iii) Llama-3.1-8B-Instruct, and (iv) Qwen3-8B. Questions are drawn from two professional-licensure bar exams: the U.S. MBE and the Brazilian OAB (Zheng et al., 2025; Calais et al., 2026). Each question is paired with a human-authored *gold passage* that justifies the correct alternative. We contrast each gold passage with an irrelevant-but-plausible *noise passage* and measure how the logit difference between the runs is recovered as a function of the patched layer and token region.

Our contributions are:

- We show that supporting evidence has a causal, not merely correlational, effect on LLM reasoning in legal QA, and we quantify this effect per layer, per token region, and per language across four open models.
- Combining the logit lens with a four-cluster question taxonomy, we find that passage integration concentrates in the upper layers (a *late-binding* signature) and is markedly asymmetric between English and Portuguese passages.

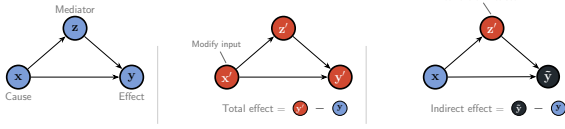


Figure 1: DAG representing interventions in activation patching. Reproduced from Mueller et al. (2024).

- We give a linguistic-structural account of why this asymmetry exists: the Portuguese OAB passages close the inferential gap from rule to fact to answer, whereas the English MBE passages require an additional bridging step that the models often fail to perform (Appendix B).

The remainder of the paper reviews activation patching, prior causal-interpretability work, and legal-reasoning evaluations of LLMs (§2), details our experimental setup (§3), and reports results (§4) before concluding with limitations and future work (§5).

2 Background and Related Work

The aim of mechanistic interpretability is to characterise *how* the internal computations of a neural network give rise to its behaviour, rather than only *what* that behaviour is (Saphra and Wiegrefe, 2024; Somvanshi et al., 2026). A growing body of work approaches this goal through causal inference, treating internal activations as variables in a latent causal system over which interventions can be performed (Geiger et al., 2021, 2024, 2025; Mueller, 2024; Mueller et al., 2024). We follow the interventionist perspective formalized by Judea Pearl (Pearl, 2001, 2012), in which structural causal models and the do operator give a precise account of what it means for one variable to *cause* another.

A structural causal model (SCM) is a tuple $\mathcal{M} = \langle \mathbf{U}, \mathbf{V}, \mathcal{F}, P(\mathbf{U}) \rangle$, where each endogenous $V_i \in \mathbf{V}$ is fixed by a structural equation $V_i := f_i(\text{Pa}_i, U_i)$ on its causal parents and an exogenous noise term, inducing an acyclic causal graph $\mathcal{G}_{\mathcal{M}}$ (Pearl, 2012; Geiger et al., 2021). Pearl’s $\text{do}(V_i = v)$ denotes a *surgical intervention*: f_i is replaced by the constant assignment $V_i := v$ while every other equation is left untouched, defining an interventional distribution $P(Y \mid \text{do}(V_i = v))$ that is generally distinct from the confounded conditional $P(Y \mid V_i = v)$ (Pearl, 2001).

In our setting the SCM is the transformer itself: $\mathbf{U}$ is the prompt, $\mathbf{V}$ are the residual-stream activations $h_t^{(\ell)}$ and output logits, and each $f_t^{(\ell)}$ is the

standard layer update, yielding the DAG of Figure 1. Activation patching is the do-operator on hidden states: with clean and corrupted caches $h_t^{(\ell),\text{clean}}$ and $h_t^{(\ell),\text{corr}}$, patching layer ℓ at token region r executes the corrupted run under

$$\text{do}\left(h_t^{(\ell)} = h_t^{(\ell),\text{clean}}\right), \quad t \in r,$$

and the recovery P_{rec} (§3) is the resulting interventional effect on the answer logit.

Identification holds because (i) *modularity* is automatic – patches overwrite activations, never weights; (ii) there is no *unobserved confounding* – the full computational graph is known; (iii) *consistency* and (iv) *non-interference* (SUTVA) follow from the TransformerLens hook mechanism and the independence of forward passes (Nanda and Bloom, 2022). The non-trivial condition is (v) *positivity*: the injected v must lie on the activation manifold, else P_{rec} measures off-distribution behaviour rather than a genuine causal effect (Heimersheim and Nanda, 2024). We enforce this by drawing every patch from a paired clean run with the same template, tokenised length, and answer slot, so $h_t^{(\ell),\text{clean}}$ is a value the network has actually computed at that (ℓ, t) .

*Activation patching*¹ operationalises this view at the level of hidden states: an activation produced by one forward pass is surgically replaced (“patched”) with the cached activation from a counterfactual pass, and the resulting change in the output logits quantifies the causal contribution of that activation (Zhang and Nanda, 2023; Heimersheim and Nanda, 2024). Figure 1 visualises the corresponding intervention graph.

Activation patching has been used to characterise how transformer LLMs perform arithmetic (Stolfo et al., 2023), how attention and MLP components route gender-bias signals (Vig et al., 2020), how multiple-choice answer selection is composed in the final layers (Wiegrefe et al., 2025), and how compositional reasoning fails in specific configurations (Li et al., 2024). These settings share crisp ground truth and short token spans. Legal QA differs in that the correct answer is mediated by domain-specific terminology, codified rules, and jurisprudence that vary across legal traditions; these properties make it an unusually demanding testbed for causal interpretability, and a setting in which

¹Activation patching is a.k.a. causal tracing or interchange intervention.

the structural form of the supporting evidence (not only its semantic relevance) plausibly matters.

Legal NLP has rapidly accumulated benchmarks aimed at evaluating LLM reasoning under realistic professional conditions, including bar admission tests, contract analysis, and case-law retrieval (Guha et al., 2023; Fei et al., 2024; Fan et al., 2025; Calais et al., 2026; Zheng et al., 2025). In parallel, hallucination-focused audits have shown that LLMs are prone to fabricating doctrine, citations, and procedural detail when not grounded in retrieved evidence (Dahl et al., 2024; Santos et al., 2025). These efforts converge on a single empirical question: do LLMs *use* legal text to reason, or do they merely surface what was memorised during pretraining (Nguyen et al., 2025)? Our work targets exactly this question at the mechanistic level.

Robust legal reasoning further requires LLMs to span diverse legal systems, integrating multilingual knowledge tailored to each system’s statutes, jurisprudence, and cultural context. This is particularly important for less-resourced jurisdictions, where the legal reality of data-rich systems risks dominating model behaviour. Our choice of U.S. MBE (common-law, English) and Brazilian OAB (civil-law, Portuguese) is deliberate, since it lets us compare causal integration across two legal traditions that differ in both language and reasoning style, exposing asymmetries that single-jurisdiction studies cannot detect.

3 Methodology

The experiment employs activation patching, a mechanistic interpretability technique, to investigate whether and how specific legal passages causally influence the reasoning process of an LLM. For each multiple-choice question, two prompts with identical structures are constructed: one contains a relevant legal passage (*gold passage*), directly related to the concept required for the correct answer, while the other contains an irrelevant passage (*noise passage*), which is legally plausible but unrelated to the specific question. The models used were Gemma-2-9B-it, Mistral-7B-v0.1, Llama-3.1-8B-Instruct, and Qwen3-8B executed via TransformerLens (Nanda and Bloom, 2022), which allows direct access to internal activations across all network layers.

Each noise passage is a human-expert-authored gold passage drawn from another question in the same examination. These passages are legally valid,

coherent, and matched in domain, register, authorship, and distributional characteristics, but are irrelevant to the evaluated question (e.g., an electoral-law passage paired with a corporate-law question). Thus, the control isolates question-specific relevance without introducing out-of-domain, shuffled, counterfactually rewritten, or model-generated text.

We select the open-weights instruction-tuned models in the 7–9B parameter range, spanning the main currently available pretraining-data regimes (predominantly English: Llama, Mistral; Chinese-augmented: Qwen; with explicit multilingual exposure: Gemma). Both datasets are chosen for being authoritative professional-licensure exams paired with human-authored explanatory passages, in two legal traditions (civil law / common law) that differ in both language and reasoning style.

The core procedure involves performing two complete forward passes, one for each prompt, and caching the residual stream activations across all layers. A systematic sweep is then conducted: for each layer, activations from the ‘*gold passage*’ run are surgically patched into the ‘*noise passage*’ run, replacing specific token blocks, after which the model completes its processing. The effect of each intervention is measured by the logit difference (LD), the difference between the logit of the correct answer and the mean logits of the incorrect alternatives at the final prompt position. Equation 1 presents the mathematical formulation of the logit difference computation. This value is normalized between 0 (corrupted run behavior) and 1 (clean gold passage behavior), producing a recovery matrix that maps, for each layer and token position, the extent to which the ‘*gold passage*’ information is causally responsible for the model’s preference for the correct answer. Figure 2 visually represents the experiment described above.

$$LD = \text{logit}(y_{\text{correct}}) - \frac{1}{|Y_{\text{wrong}}|} \sum_{y \in Y_{\text{wrong}}} \text{logit}(y) \quad (1)$$

We calculated LD twice: once with the gold passage (LD_{clean}) and once with the noise passage (LD_{corr}). The difference between them is as follows:

$$\Delta = LD_{\text{clean}} - LD_{\text{corr}} \quad (2)$$

For each layer l and region r , we compute a patched logit difference (LD_{patched}) by replacing the residual stream activations at region r and layer l in the corrupted run with the corresponding activations from the clean run, then completing the

VISUAL REPRESENTATION OF THE EXPERIMENT: ACTIVATION PATCHING FOR LEGAL REASONING

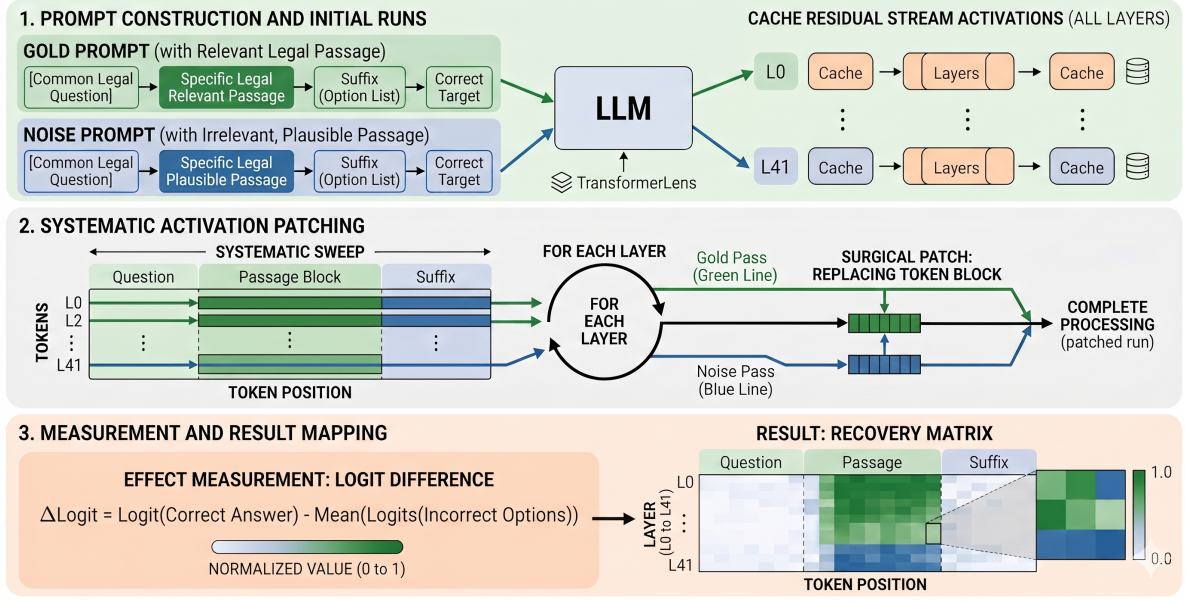


Figure 2: Visual representation of Activation Patching experiment used in this paper

forward pass. Normalized recovery (P_{rec}) is defined as follows:

$$P_{\text{rec}} = \frac{LD_{\text{patched}} - LD_{\text{corr}}}{LD_{\text{clean}} - LD_{\text{corr}}} = \frac{LD_{\text{patched}} - LD_{\text{corr}}}{\Delta} \quad (3)$$

yielding 0 when the patch has no effect and 1 when it fully restores the clean-run behaviour, thereby quantifying the causal contribution of each (layer, region) pair to the difference in model behaviour induced by the passage substitution.

To obtain multiple-choice questions from the legal domain, we utilized two available datasets consisting of examination questions for professional legal licensure in Brazil (OAB test) and the United States (MBE test). Both datasets, compiled by Calais et al. (2026) and Zheng et al. (2025), aim to enable the evaluation of legal reasoning in LLMs and performance enhancement via techniques such as RAG. Each dataset includes the question prompts, alternatives, the expected correct answer, and a *gold passage*, a human-written text that justifies, exemplifies, or outlines the guidelines and reasoning demonstrating the validity of the correct alternative. Although the textual composition varies between the two datasets, the purpose of this field remains the same: to provide textual evidence supporting the correct answer. We refer to this human-written text as the *supporting evidence*, since it represents external information (e.g., relevant case law) intended to guide the model toward

the appropriate response.

This human-written text for each question differs from what might traditionally be termed *context*. Based on the text typology developed by Halliday and Matthiessen (Halliday and Matthiessen, 2013; Matthiessen and Teruya, 2023), which classifies texts according to their purposes, we would categorize this text as the *field of experience*. This is because it does not merely explain the prompt or the alternatives; rather, it provides a textual explanation of the cultural sphere in which the question is situated (e.g., Constitutional Law, Tax Law, or International Law). It is explanatory and argumentative in nature, validating the human reasoning that culminates in the correct alternative.

4 Results and discussion

4.1 Accuracy and logit difference

As shown in Table 1, the models vary substantially in their accuracy across the evaluated datasets. Llama-3.1-8b-Instruct achieves the best overall performance (considering both the Brazilian and U.S. datasets), with an accuracy of 0.82 when provided with the gold passage and the smallest performance gap relative to the noise condition (0.24). This may suggest that the model’s advantage over the others is not based exclusively on supporting evidence, but could also come from characteristics of its pretraining data.

When examining performance under the gold

Model	Gold	Noise	Gold EN	Gold PT	Noise EN	Noise PT
Llama-3.1-8b-Instruct	0.82	0.58	0.74	0.90	0.63	0.52
Gemma-2-9b-it	0.74	0.45	0.59	0.86	0.41	0.48
Mistral-7b-v0.1	0.62	0.48	0.68	0.57	0.42	0.52
Qwen3-8b	0.38	0.40	0.42	0.33	0.47	0.33

Table 1: Overall and language-specific accuracy across gold and noisy conditions.

passage condition, Gemma-2-9b-it exhibits the greatest asymmetry between the two languages analysed (0.86 vs. 0.59), showing substantially better performance in Portuguese. This is an initially unexpected effect, given that the model was most likely trained on a considerably larger amount of English-language data compared to Brazilian Portuguese data. In contrast, Mistral-7B shows better performance with gold passages in English than in Portuguese (0.67 vs. 0.57), most likely reflecting the predominance of English-language data during pretraining.

However, when we analyse the causal influence of the passage (Table 2) through the $|\Delta|$ metric described in Section 3, we observe that the largest measurable causal effect of the passage occurs in Gemma-2-9b-it. The gold passage shifts the model toward the correct alternative far more abruptly than any other model evaluated in this study. This effect is even stronger in Portuguese (3.699 vs. 2.613 in English), which could suggest that the gold passages in the Brazilian dataset generate stronger causal signals than those in the U.S. dataset.

Even when the models answer a dataset question incorrectly, however, a positive LD still reveals an interesting finding regarding the causal influence of the gold passages. In such cases, we have:

$$\text{logit}(y_c) > \text{mean}(\text{logit}(y_i))$$

and simultaneously:

$$\text{logit}(y_x) > \text{logit}(y_c)$$

where y_c denotes the correct alternative, y_i the incorrect alternatives, and y_x the highest-scoring competing alternative.

This phenomenon occurs when a specific alternative receives a logit value sufficiently high to surpass the correct answer individually, even though the remaining incorrect alternatives have very low logits, thereby lowering the overall mean. In this scenario, the gold passage raises the logit of the correct answer above the average of the field, but not

above that of the strongest competing alternative. In other words, the gold passage is causally steering the model in the correct direction, but the shift is insufficient to overcome the dominant competitor.

Model	$ \Delta $	$ \Delta $ EN	$ \Delta $ PT
Gemma-2-9b-it	3.213	2.613	3.699
Llama-3.1-8B-it	1.806	1.272	2.290
Mistral-7B-v0.1	0.736	0.923	0.566
Qwen3-8B	0.293	0.298	0.288

Table 2: Absolute $|\Delta|$ values across models and languages.

These results indicate that a positive clean LD associated with an incorrect answer should not be interpreted as a failure of the passage; rather, it constitutes evidence that the passage is functioning partially as intended. Without the passage, the model would likely exhibit an even more negative LD and would therefore be even farther from the correct answer. The gold passage produces a stronger causal shift toward the correct alternative than the noise passage, even in cases where the model ultimately fails under both conditions.

4.2 Logit lens and late binding

Another relevant question concerns the layer at which the model begins to exhibit a substantially stronger preference for the correct alternative. We further ask whether this preference remains consistent across layers and prompt-token regions (question, answer alternatives, passages), and whether causal patching interventions influence this process. For each question in the datasets, we additionally employed the logit lens technique (nostalgebraist, 2020) in order to observe the progression of the models’ reasoning throughout the forward pass.

For the Gemma-2-9b-it and Llama-3.1-8B-Instruct models, we observed a result consistent with what the mechanistic interpretability literature has referred to as late binding: for many transformer tasks, the model’s output-aligned decision, that is, its next-token or

label preference, often stabilizes only in the upper layers, frequently within the last quarter of the network depth or in the final blocks rather than exclusively in the last layer itself (Geva et al., 2021; Belrose et al., 2023; Halawi et al., 2024; Li et al., 2024).

For questions that the models answered correctly, we observed a consistent pattern (Figure 3) in which, throughout approximately 80% of them, the models did not appear to make any explicit prediction or commitment regarding the correct alternative. Instead, the intermediate layers seemed primarily devoted to processing or reasoning about the legal case presented in the prompt. Only in the upper layers did the models return to the final objective of selecting one of the alternatives, subsequently maintaining this choice consistently and with high probability until the end of the forward pass.

For questions answered incorrectly, we similarly observed that the models tended to defer commitment to a specific alternative until the upper layers. However, in these cases, the probability assigned to the correct alternative remained consistently very low, often effectively zero, without the model ever substantially considering competing alternatives, both in the clean and corrupted runs, including in the final layers. Even when provided with the gold passage, once the model converged toward an incorrect alternative after the reasoning process developed across approximately 80% of its layers, it tended to remain committed to the error, without reconsidering alternative options in the final layers or exhibiting any abrupt late-stage shift in preference.

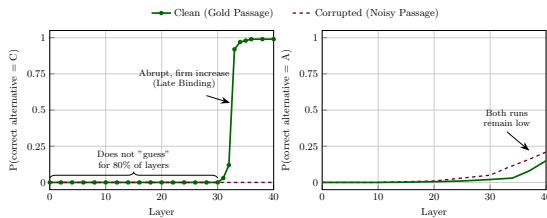


Figure 3: Logit lens analysis on legal multiple-choice QA. Left: Correct answer. Right: Incorrect answer. Solid line = gold passage; dashed = noise passage.

4.3 Passage recovery (P_{rec}) and questions clusters

Passage recovery quantifies the causal contribution of the passage region to the difference in model

behaviour between the clean and corrupted runs. High recovery in correct answers combined with low or negative recovery in incorrect ones indicates that the model is integrating passage content in a way that is causally decisive for the right answer, not merely processing it syntactically. When recovery is similarly high in both outcomes, or inverted, the model absorbs the passage as a contextual signal without reliably completing the inferential step.

In our study (Table 3), Gemma-2-9b-it shows the clearest language asymmetry. In Portuguese, the gap between correct and incorrect answers (+0.144, rising to +0.315 when only high- Δ cases are considered) indicates that passage integration is consistently causal for success and largely absent in failure. In English, the gap almost vanishes (+0.047), with incorrect answers reaching 0.453, nearly identical to correct ones at 0.500, suggesting that English common-law passages activate internal representations without reliably completing the inferential chain. Llama-3.1-8B-Instruct presents the strongest overall separation, particularly in Portuguese (+0.599), where incorrect answers show negative recovery (-0.126), indicating the passage is actively absent from the causal pathway in failure cases.

Mistral-7B-v0.1 and Qwen3-8B exhibit the most divergent patterns. Mistral’s English inversion (gap = -0.634) largely reflects noise amplification from its high rate of near-zero Δ cases, while its Portuguese results remain coherent (+0.267). Qwen3-8B, by contrast, shows a structurally inverted pattern in English even under stable conditions: incorrect answers carry higher passage recovery (0.555) than correct ones (0.125), with a gap of -0.430. This suggests the model is extracting passage-level signals but routing them toward wrong answers, a mechanistic failure mode distinct from simple passage blindness.

The cluster taxonomy introduced in this work provides an empirically grounded framework for characterising how language models interact with retrieved legal passages. Each question is assigned to one of four behavioural categories based on the model’s top answer under both the gold and noise passage conditions: Cluster A (correct under both), Cluster B (correct only with the gold passage), Cluster C (incorrect under both), and Cluster D (gold passage actively worsens performance). This classification emerges directly from the experimental observations rather than from any prior theoretical assumption, and its discriminative power is evi-

Model	Lang	Correct	Incorrect	Gap
Llama-3.1-8B-Instruct	ALL	0.398	0.137	+0.261
	EN	0.295	0.241	+0.053
	PT	0.474	-0.126	+0.599
Gemma-2-9b-it	ALL	0.462	0.406	+0.056
	EN	0.500	0.453	+0.047
	PT	0.440	0.296	+0.144
Mistral-7B-v0.1	ALL	0.167	0.306	-0.139
	EN	-0.016	0.617	-0.634
	PT	0.365	0.098	+0.267
Qwen3-8B	ALL	0.272	0.470	-0.198
	EN	0.125	0.555	-0.430
	PT	0.440	0.402	+0.038

Table 3: Mean passage recovery for correct and incorrect answers (gold passage condition). Correct and Incorrect columns report the mean normalised LD recovery of the passage block for questions answered correctly and incorrectly, respectively. Gap = Correct - Incorrect.

denced by the systematic variation it reveals across models and languages. Notably, the clusters admit a mechanistic reading: Cluster B isolates questions where the passage is causally necessary for the correct answer, Cluster C identifies the limits of passage-based reasoning, and Cluster D exposes the subset of cases where retrieval augmentation is actively harmful.

The distribution across clusters (reported in full in Appendix A) reveals several findings with direct implications for reasoning interpretation. Llama-3.1-8B-Instruct achieves the highest combined A+B rate (82%), indicating that it either already knows the answer or successfully uses the passage to reach it, with only 10% of questions falling into Cluster C. Qwen3-8B is the only model with 0% Cluster B across all language conditions, indicating that its passage integration never functions as a decisive causal trigger: the model either answers independently of the passage or fails regardless of it. Gemma-2-9b-it and Llama-3.1-8B-Instruct both exhibit 0% Cluster D in Portuguese, meaning the gold passage never degrades performance on Brazilian legal questions for either model, while English questions yield non-trivial Cluster D rates for all models, a result that aligns with the passage recovery asymmetry observed between languages.

4.4 Pretraining and legal semantic specificities

Among the various QA results obtained, particularly noteworthy is the case concerning the legal division of assets in the stable union ("*união estável*") of Gloria and Paulo, under Brazilian civil law. When queried about the correct answer for this case, the Llama, Mistral, and Gemma models all selected the incorrect alternative, even when provided with a detailed legal explanation of the

case through the gold passage.

This error may be associated with insufficient prior training on data pertaining to Brazilian Family Law and Succession Law. An additional factor that may account for this failure is these models' limited semantic comprehension of certain expressions characteristic of the legal context.

A salient example is the verb *to compete* (*competir*) in the domain of Succession Law. Rather than involving literal interpretation, the challenge stems from the polysemy of the verb *competir*, which, in legal contexts, may denote the entitlement or allocation of assets to lawful heirs. This usage is considerably less frequent and more domain-restricted than its common association with rivalry or competition, potentially leading language models to favor the more prototypical sense. Notably, the capacity for semantic comprehension of legal terminology is among the core competencies expected of a law school graduate sitting for the national bar examination of the Brazilian Bar Association.

These observations demonstrate that, depending on the domain of knowledge, in order for AI models to achieve greater accuracy and quality in their responses, gold passages must encompass not only the technical and contextual information pertaining to a given subject, but also linguistic and terminological guidance, that is, explicit instruction regarding the meaning of nouns, verbs, and expressions that are characteristic of the relevant domain (Jayakumar et al., 2023; Breton et al., 2025; Siino et al., 2025).

4.5 Controlling for language effects

To disentangle language from dataset-specific and legal-tradition effects, we conducted a 2×2 follow-up crossing language (Portuguese/English) with

legal tradition (Brazilian civil law/OAB and U.S. common law/MBE). Each dataset was evaluated in its original language and in translation, allowing language to vary while question content and passage structure remained fixed. The questions, answer options, gold passages, and noise passages were automatically translated using GPT-5.5 High, which provides high translation accuracy while preserving the original semantic content (Van Doren et al., 2026).

As shown in Figure 4, translating OAB into English generally preserved or improved accuracy, whereas translating MBE into Portuguese reduced accuracy for most models; conversely, differences between OAB and MBE within the same language were particularly large for Gemma and Llama. The mechanistic results show an even clearer pattern (Figure 5): $|\Delta|$ is consistently larger for OAB than MBE in both languages across all three models analyzed, while translation increases $|\Delta|$ for OAB and decreases it for MBE. These results suggest that the observed asymmetry is more strongly associated with legal-domain/discourse properties than with language alone, although differences in multilingual pretraining exposure and translation remain potential confounds.

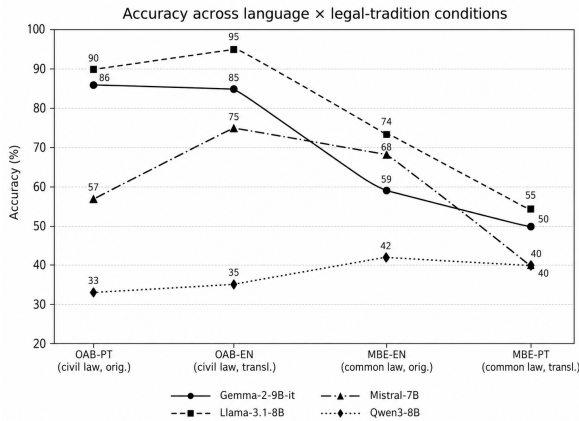


Figure 4: Accuracy across the 2×2 language-legal tradition control. Translating OAB into English largely preserves or improves performance, whereas translating MBE into Portuguese generally reduces it.

4.6 Discourse structure and causal passage utility

Passage effectiveness depends not only on semantic relevance, but on how discourse relations organize the relevant information and how explicitly the text supplies the inferential link between evidence and conclusion. This view is consistent with

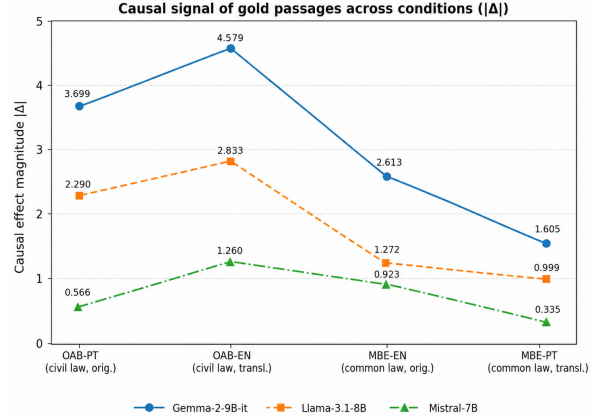


Figure 5: Causal effect magnitude ($|\Delta|$) across language-legal tradition conditions.

discourse theories in which coherence relations structure how propositions are connected and interpreted, while implicit relations require additional inferential work from the comprehender (Mann and Thompson, 1988; Kintsch, 1988; Sanders et al., 1992; Sanders and Noordman, 2000). Accordingly, the Portuguese and English gold passages differ not only in content, but also in how directly their discourse organization supports the target answer.

Portuguese passages typically realize an answer-directed chain in which a legal rule is explicitly applied to case-specific facts (*rule* $\rightarrow$ *fact* $\rightarrow$ *answer*), thereby reducing the inferential gap between evidence and response. English passages more often elaborate doctrine through definitions and precedent (*rule* $\rightarrow$ *rule chains of citations*), leaving its application to the question facts as an implicit bridging inference. This distinction provides a discourse-theoretic account of our causal results: relevant evidence may be available to the model without becoming causally decisive when the passage does not explicitly encode the relation linking that evidence to the answer.

5 Conclusion

We used activation patching to test whether human-authored supporting evidence causally shapes LLM behaviour on legal multiple-choice QA. Three findings stand out. First, gold passages exert a measurable causal effect on predictions, with absolute logit-difference shifts ranging from $|\Delta| = 0.293$ for Qwen3-8B to $|\Delta| = 3.213$ for Gemma-2-9b-it (Table 2); the effect is therefore model-dependent and far from uniform. Second, logit-lens analysis on Gemma-2-9b-it and Llama-3.1-8B-Instruct

shows that the answer-aligned decision concentrates in the upper layers (the late-binding regime), and that incorrect answers exhibit no late-stage reconsideration even when the gold passage is provided. Third, the four-cluster taxonomy reveals a strong language asymmetry: the gold passage never degrades performance on Portuguese OAB questions for Llama or Gemma (0% Cluster D), while English MBE questions yield non-trivial Cluster D rates across every model.

These findings have implications for two audiences. For mechanistic interpretability, legal QA inherits the late-binding regime previously observed in arithmetic and multiple-choice tasks, but with a sharper dependence on the *form* of the textual input than those settings show – suggesting that the locus of integration in upper layers is modulated by discourse structure, not only by semantic content. For legal NLP and retrieval-augmented systems, our results indicate that passage relevance is not sufficient to predict whether a passage helps: the structural shape of the passage (*rule* $\rightarrow$ *fact* $\rightarrow$ *answer* vs. *rule* $\rightarrow$ *rule chains of citations*, Appendix B) mediates whether the model integrates it causally.

For many questions, particularly those belonging to Cluster A (correct answers under both the gold passage and noise passage conditions), there appears to be a strong influence from pre-training and/or instruction tuning directed toward question-answering tasks in models such as Llama-3.1-8B-Instruct and Gemma-2-9B-it. This training likely included examples of legal reasoning and contextualized QA tasks, thereby reinforcing both internal knowledge representations (Cluster A) and the capacity to effectively use passages when necessary (Cluster B).

Our experiment cannot fully disentangle internally acquired knowledge from the influence exerted by the gold passage when the models answer correctly. What it does more clearly reveal, among the analysed models, is that Llama-3.1-8B-Instruct demonstrates the best balance between internal knowledge and the ability to leverage passages: neither as dependent on the passage as Gemma-2-9B-it in Portuguese, whose performance drops substantially without the gold passage, nor as indifferent to the passage as Qwen3-8B.

The Portuguese dataset, by contrast with the English passages described in Sec. B, attempts to reconcile the legal situation presented in the question with the relevant jurisprudence, explicitly describ-

ing the argumentative and often inferential reasoning steps that legal professionals would be expected to perform in order to correctly answer the question. In many cases where the high-performing models fail even with Portuguese gold passages, the errors appear to stem from highly domain-specific linguistic knowledge, such as the use of legal expressions, lexical items, or verbs unknown to the model and not verbally explicated within the gold passages themselves. As a result, the models fail to capture these implicit linguistic cues and ultimately produce incorrect answers. A comparison between the linguistic structures of the Portuguese and English gold passages is provided in Appendix B.

Promising directions include applying sparse autoencoders (Cunningham et al., 2023; Makelov et al., 2025) to identify the specific features that mediate passage integration, comparing instruction-tuned and base models to isolate the contribution of supervised fine-tuning to the late-binding effect, and extending the study to additional jurisdictions and matched passage forms across languages so that linguistic structure can be varied independently from legal content.

Limitations

We evaluate four open-weights instruction-tuned models in the 7–9B parameter range on two professional-licensure bar exams. Larger open models, base (non-instruction-tuned) models, and closed-source models such as GPT-5.5 are out of scope; a more systematic comparison between instruction-tuned and base variants, as partially observed with Qwen3-8B, would be particularly informative for the late-binding effect we report.

Activation patching provides evidence of necessity or sufficiency of specific activations under the specific intervention performed; it does not establish that the model’s full reasoning chain is captured by the patched regions, nor that the same components would mediate behaviour under different prompting or sampling regimes.

Our two datasets correlate language with legal tradition (Portuguese–civil law, English–common law). The asymmetric passage integration we observe could be driven by language, by reasoning style, by passage structure (Appendix B), or by a combination of these factors. Disentangling them would require parallel corpora across traditions, or paired passages with matched discourse structure across languages, a direction we leave to future

work.

All reported Δ and P_{rec} values are single-pass deterministic estimates (Sec.3). Variability across questions is reflected in the cluster-level analysis, but variability across model revisions, prompt formats, and floating-point non-determinism is not quantified.

Future work should expand to additional datasets, languages, and jurisdictions, employing gold passages that are more textually aligned across languages so that textual quality, discourse organization, and linguistic structure can be varied independently of legal content.

Acknowledgments

This work was supported by CI-IA Saúde, CNPq, CAPES, FAPEMIG, FAPESP, IAIA-INCT on AI and INCT-TILDIAR.

References

- Nora Belrose, Igor Ostrovsky, Lev McKinney, Zach Furman, Logan Smith, Danny Halawi, Stella Biderman, and Jacob Steinhardt. 2023. Eliciting latent predictions from transformers with the tuned lens. *arXiv preprint arXiv:2303.08112*.
- Julien Breton, Mokhtar Mokhtar Billami, Max Chevalier, Ha Thanh Nguyen, Ken Satoh, Cassia Trojahn, and May Myo Zin. 2025. Leveraging LLMs for legal terms extraction with limited annotated data. *Artificial Intelligence and Law*, pages 1–27.
- Pedro Calais, Janderson Santos, Anisio Lacerda, and Wagner Meira Jr. 2026. Monotonic scaffolding as a diagnostic lens for legal reasoning in LLMs. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (ACL 2026)*, San Diego, California. Association for Computational Linguistics.
- Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. 2023. Sparse autoencoders find highly interpretable features in language models. *arXiv preprint arXiv:2309.08600*.
- Matthew Dahl, Varun Magesh, Mirac Suzgun, and Daniel E Ho. 2024. Large legal fictions: Profiling legal hallucinations in large language models. *Journal of Legal Analysis*, 16(1):64–93.
- Yu Fan, Jingwei Ni, Jakob Merane, Yang Tian, Yoan Hermstrüwer, Yinya Huang, Mubashara Akhtar, Etienne Salimbeni, Florian Geering, Oliver Dreyer, and 1 others. 2025. Lexam: Benchmarking legal reasoning on 340 law exams. *arXiv preprint arXiv:2505.12864*.
- Zhiwei Fei, Xiaoyu Shen, Dawei Zhu, Fengzhe Zhou, Zhuo Han, Alan Huang, Songyang Zhang, Kai Chen, Zhixin Yin, Zongwen Shen, and 1 others. 2024. Law-bench: Benchmarking legal knowledge of large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7933–7962.
- Atticus Geiger, Duligur Ibeling, Amir Zur, Maheep Chaudhary, Sonakshi Chauhan, Jing Huang, Aryaman Arora, Zhengxuan Wu, Noah Goodman, Christopher Potts, and 1 others. 2025. Causal abstraction: A theoretical foundation for mechanistic interpretability. *Journal of Machine Learning Research*, 26(83):1–64.
- Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. 2021. Causal abstractions of neural networks. *Advances in Neural Information Processing Systems*, 34:9574–9586.
- Atticus Geiger, Zhengxuan Wu, Christopher Potts, Thomas Icard, and Noah Goodman. 2024. Finding alignments between interpretable causal variables and distributed neural representations. In *Causal Learning and Reasoning*, pages 160–187. PMLR.
- Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2021. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5484–5495.
- Maxime Griot, Coralie Hemptinne, Jean Vanderdonckt, and Demet Yuksel. 2025. Large language models lack essential metacognition for reliable medical reasoning. *Nature Communications*, 16(1):642.
- Neel Guha, Julian Nyarko, Daniel Ho, Christopher Ré, Adam Chilton, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel Rockmore, Diego Zambano, and 1 others. 2023. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. *Advances in Neural Information Processing Systems*, 36:44123–44279.
- Danny Halawi, Jean-Stanislas Denain, and Jacob Steinhardt. 2024. Overthinking the truth: Understanding how language models process false demonstrations. In *International Conference on Learning Representations*, volume 2024, pages 42749–42787.
- Michael Halliday and Christian Matthiessen. 2013. *Halliday’s introduction to functional grammar*. Routledge.
- Stefan Heimersheim and Neel Nanda. 2024. How to use and interpret activation patching. *arXiv preprint arXiv:2404.15255*.
- Thanmay Jayakumar, Fauzan Farooqui, and Luqman Farooqui. 2023. Large language models are legal but they are not: Making the case for a powerful legal-llm. In *Proceedings of the Natural Legal Language Processing Workshop 2023*, pages 223–229.
- Manuj Kant, Sareh Nabi, Manav Kant, Roland Scharrer, Megan Ma, and Marzieh Nabi. 2025. Towards robust legal reasoning: Harnessing logical LLMs in law. *arXiv preprint arXiv:2502.17638*.

- Walter Kintsch. 1988. The role of knowledge in discourse comprehension: a construction-integration model. *Psychological review*, 95(2):163.
- Zhaoyi Li, Gangwei Jiang, Hong Xie, Linqi Song, Defu Lian, and Ying Wei. 2024. Understanding and patching compositional reasoning in LLMs. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9668–9688.
- Aleksandar Makelov, Georg Lange, and Neel Nanda. 2025. Towards principled evaluations of sparse autoencoders for interpretability and control. In *International Conference on Learning Representations*, volume 2025, pages 33588–33636.
- William C. Mann and Sandra A. Thompson. 1988. *Rhetorical structure theory: Toward a functional theory of text organization*. *Text - Interdisciplinary Journal for the Study of Discourse*, 8(3):243–281.
- Christian MIM Matthiessen and Kazuhiro Teruya. 2023. *Systemic functional linguistics: A complete guide*. Routledge.
- Aaron Mueller. 2024. Missed causes and ambiguous effects: Counterfactuals pose challenges for interpreting neural networks. In *ICML 2024 Workshop on Mechanistic Interpretability*.
- Aaron Mueller, Jannik Brinkmann, Millicent Li, Samuel Marks, Koyena Pal, Nikhil Prakash, Can Rager, Aruna Sankaranarayanan, Arnab Sen Sharma, Jiding Sun, and 1 others. 2024. The quest for the right mediator: Surveying mechanistic interpretability through the lens of causal mediation analysis. *arXiv preprint arXiv:2408.01416*.
- Neel Nanda and Joseph Bloom. 2022. Transformerlens. <https://github.com/TransformerLensOrg/TransformerLens>.
- Ha Thanh Nguyen, Wachara Fungwacharakorn, May Myo Zin, Randy Goebel, Francesca Toni, Kostas Stathis, and Ken Satoh. 2025. LLMs for legal reasoning: A unified framework and future perspectives. *Computer Law & Security Review*, 58:106165.
- nostalgebraist. 2020. Interpreting gpt: The logit lens. <https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens>. Accessed: 2026-05-20.
- Judea Pearl. 2001. Direct and indirect effects. In *Probabilistic and causal inference: the works of Judea Pearl*, pages 373–392. Association for Computing Machinery, New York, NY, USA.
- Judea Pearl. 2012. The causal mediation formula—a guide to the assessment of pathways and mechanisms. *Prevention Science*, 13(4):426–436.
- Ted JM Sanders and Leo GM Noordman. 2000. The role of coherence relations and their linguistic markers in text processing. *Discourse processes*, 29(1):37–60.
- Ted JM Sanders, Wilbert PM Spooren, and Leo GM Noordman. 1992. Toward a taxonomy of coherence relations. *Discourse processes*, 15(1):1–35.
- Janderson Santos, Pedro Calais, Ricardo Shen, Gabriel Rolla, Ana Luísa Araújo, Fernanda Carvalho, Victor Hon Fonseca, Aziz Saliba, Anisio Lacerda, and Wagner Meira Jr. 2025. Avaliação do raciocínio jurídico em modelos de linguagem com base na estrutura IRAC. In *Anais do XVII Congresso Brasileiro de Inteligência Computacional (CBIC 2025)*, Belo Horizonte, Brasil. Sociedade Brasileira de Inteligência Computacional.
- Naomi Saphra and Sarah Wiegrefe. 2024. Mechanistic? In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 480–498.
- Marco Siino, Mariana Falco, Daniele Croce, and Paolo Rosso. 2025. Exploring LLMs applications in law: A literature review on current legal NLP approaches. *IEEE Access*, 13:18253–18276.
- Shriyank Somvanshi, Md Monzurul Islam, Amir Rafe, Anannya Ghosh Tusti, Arka Chakraborty, Anika Baitullah, Tausif Islam Chowdhury, Nawaf Alnawmasi, Anandi Dutta, and Subasish Das. 2026. Bridging the black box: A survey on mechanistic interpretability in ai. *ACM Computing Surveys*, 58(8):1–35.
- Alessandro Stolfo, Yonatan Belinkov, and Mrinmaya Sachan. 2023. A mechanistic interpretation of arithmetic reasoning in language models using causal mediation analysis. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7035–7052.
- Madison Van Doren, Casey Ford, Jennifer Barajas, Riley VanMeter, and Cory Holland. 2026. “be my cheese?”: Cultural nuance benchmarking for machine translation in multilingual llms. In *Proceedings of the Fifth Workshop on Generation, Evaluation and Metrics (GEM)*, pages 52–76.
- Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. 2020. Investigating gender bias in language models using causal mediation analysis. *Advances in Neural Information Processing Systems*, 33:12388–12401.
- Sarah Wiegrefe, Oyvind Tafjord, Yonatan Belinkov, Hanna Hajishirzi, and Ashish Sabharwal. 2025. Answer, assemble, ace: Understanding how llms answer multiple choice questions. In *International Conference on Learning Representations*, volume 2025, pages 78293–78318.
- Fred Zhang and Neel Nanda. 2023. Towards best practices of activation patching in language models: Metrics and methods. *arXiv preprint arXiv:2309.16042*.
- Lucia Zheng, Neel Guha, Javokhir Arifov, Sarah Zhang, Michal Skreta, Christopher D Manning, Peter Henderson, and Daniel E Ho. 2025. A reasoning-focused

legal retrieval benchmark. In *Proceedings of the 2025 Symposium on Computer Science and Law*, pages 169–193.

A Appendix A: Question Clusters

Table 4 reports the full distribution across the four behavioral clusters introduced in Sec.4.

B Appendix B: Linguistic and Structural Analysis of Gold Passages

This appendix examines the compositional differences between the English and Portuguese gold passages used in the activation patching experiments. Although both passage types are designed to justify a single correct answer choice, they differ substantially in structure, reasoning style, and the degree to which they bridge legal doctrine to the specific question facts. We argue that these differences contribute to the asymmetric passage integration observed across models (Section 4).

B.1 Annotated passage examples

Labels:

- Legal rule / principle
- Application to case facts
- Calculation / conclusion
- Citation / external authority
- Exclusion / limitation

Passage in Portuguese:

(*uniao_estavel_sucessao_ascendentes*)

Considerando que Paulo e Glória viviam em união estável sem pacto escrito,

o regime de bens aplicável é o da comunhão parcial. Assim, separa-se a meação do patrimônio comum, que é de direito próprio da companheira.

O patrimônio total de R\$ 1.600.000,00 será dividido, cabendo R\$ 800.000,00 à companheira a título de meação. O valor restante, R\$ 800.000,00, constitui a herança de Paulo.

Na sucessão, a companheira sobrevivente é equiparada ao cônjuge. Conforme a lei civil, o companheiro sobrevivente concorre com os ascendentes na herança.

Paulo não deixou descendentes nem pais, mas deixou avós, que são ascendentes de segundo grau.

A companheira recebe metade da herança e a outra metade é destinada aos ascendentes. Portanto, a companheira herdará R\$ 400.000,00; os quatro avós dividirão os R\$ 400.000,00 restantes.

Os irmãos de Paulo, sendo herdeiros colaterais, são excluídos da sucessão na presença de ascendentes e da companheira.

Passage in English:

(*mortgagee_in_possession_tort_liability*)

A mortgagee may become a mortgagee in possession where the mortgage itself gives the mortgagee the right to enter and “take possession of the mortgaged premises and receive the rents and profits”

(*Gomez v. Bobker*, 124 A.D.2d 703, 704, 508 N.Y.S.2d 215; *Gasco Corp. & Gordian Group of Hong Kong v. Tosco Props.*, 236 A.D.2d 510, 653 N.Y.S.2d 687).

A mortgagee in possession “takes the rents and profits in the quasi character of trustee or bailiff of the mortgagor”

(*Gasco Corp.*, 236 A.D.2d at 512, 653 N.Y.S.2d 687).

“[A] mortgagee cannot be in possession without consent of the mortgagor”

(*Hoye v. Theresa*, 134 App.Div. 255, 256, 118 N.Y.S. 951; *Gordon v. Krellman*, 207 App.Div. 773, 202 N.Y.S. 682).

“In order to acquire the rights of a mortgagee in possession it [is] necessary for the [party claiming such status] to show that it had entered with the mortgagor’s consent, or had otherwise entered lawfully”

(*Herrmann v. Cabinet Land Co.*, 217 N.Y. 526, 528, 112 N.E. 476).

B.2 Structural Comparison

B.3 Implications for Passage Integration

The Portuguese passage applies a legal rule directly to the case facts and derives a numerical answer, effectively closing the inferential gap between passage and correct answer choice. The model need

Model	Lang	A	B	C	D
Llama-3.1-8B-Instruct	ALL	50%	32%	10%	8%
	EN	47%	26%	11%	16%
	PT	52%	38%	10%	0%
Gemma-2-9b-it	ALL	42%	32%	23%	3%
	EN	35%	24%	35%	6%
	PT	48%	38%	14%	0%
Mistral-7B-v0.1	ALL	40%	22%	30%	8%
	EN	37%	32%	26%	5%
	PT	43%	14%	33%	10%
Qwen3-8B	ALL	38%	0%	60%	2%
	EN	42%	0%	53%	5%
	PT	33%	0%	67%	0%

Table 4: Cluster distribution by model and language. **A**: correct under both gold and noise passages. **B**: correct only with the gold passage. **C**: incorrect under both passages. **D**: gold passage degrades performance relative to noise.

Feature	Portuguese (PT)	English (EN)
Reasoning style	Deductive / applied	Precedent-based
Case citations	None	5+
Fact-specific content	Yes (names, amounts)	No
Inference chain	Rule $\rightarrow$ fact $\rightarrow$ answer	Rule $\rightarrow$ rule
Dominant component	Application + calculation	Citation + definition
Gap to answer	Closed by passage	Requires extra step

Table 5: Structural features of Portuguese and English gold passages.

only follow the chain already made explicit in the text.

The English passage, by contrast, defines and delimits a legal doctrine through a sequence of case citations, without connecting that doctrine to the specific facts of the question. The model must independently perform the bridging step, recognising that the described doctrine applies to the scenario and selecting the answer that reflects it.

This structural difference provides a plausible account of the language asymmetry in passage recovery reported in Table 3: when the passage delivers a completed argument, integration is more likely to be causally decisive; when it delivers only doctrinal background, integration may occur without producing the correct inferential output.

C Appendix C: Experimental Setup

All runs used deterministic greedy decoding: temperature=0, with top-p and top-k sampling disabled. Inference on GPU used bfloat16 (NVIDIA A100, 80GB); float16 was the CPU fallback. Models were loaded without quantisation via TransformerLens (v3.0.0) with fold_ln=False,

center_writing_weights=False, and center_unembed=False to preserve the residual stream activation geometry required for valid patching. Activation patching targeted the residual stream at hook_resid_pre for each layer; the logit lens used hook_resid_post projected through the unembedding matrix. When gold and noise passages differ in token length, clean activations were resampled to the corrupted sequence length via linear interpolation along the sequence axis prior to substitution. All results are exactly reproducible given identical model weights, as no stochastic component is present in any stage of the pipeline.