Monotonic Scaffolding as a Diagnostic Lens for Legal Reasoning in LLMs

Pedro Calais
UFMG, Brazil
pcalais@dcc.ufmg.br

Janderson Santos
UFMG, Brazil
jandersonsantos@dcc.ufmg.br

Anisio Lacerda
UFMG, Brazil
anisio@dcc.ufmg.br

Wagner Meira Jr.
UFMG, Brazil
meira@dcc.ufmg.br

Abstract

Modern evaluation of Legal QA systems is shifting from terminal accuracy toward process-aware analyses of model reasoning. We propose a diagnostic framework grounded in monotonic pedagogical scaffolding, where language models receive gold-standard, case-relevant information across stages aligned with the canonical legal framework FIRAC — Facts, Issue, Rules, Application, Conclusion. By strictly adding solution-relevant content at each step, we introduce a controlled monotonic intervention that allows for the evaluation of reasoning trajectories rather than isolated outcomes.

This longitudinal design enables the introduction of two transition-based diagnostics: Errors-to-Success (E2S) quantifies the guidance required to reach correctness, while Success-to-Errors (S2E) measures the fragility of that correctness under additional structure. These local patterns define a global robustness criterion termed Stable Accuracy, which credits a response only if the model maintains correctness throughout all scaffolding stages and enforces a higher bar for correctness by distinguishing sustained reasoning from transient patterns.

We instantiate the framework on 3,123 Brazilian Bar Exam questions paired with expert-annotated explanations. Our findings reveal model instability patterns hidden from accuracy-only metrics and demonstrate that terminal accuracy systematically overestimates legal reasoning competence. To test the robustness of our diagnostics, we also evaluate a majority-vote aggregation across multiple reasoning samples, finding that the observed instability patterns persist under this stronger inference setting. Furthermore, principal component analysis indicates that legal domains cluster into distinct regions, suggesting systematic differences in reasoning demands across domains. While focused on the legal domain, our evaluation protocol is generalizable to any task with a staged reasoning structure.

1 Introduction

Large language models (LLMs) are becoming increasingly central to the legal domain (Wu et al., 2024; Dehghani et al., 2025). Recent systems have demonstrated strong performance across core legal NLP tasks—such as legal question answering, case retrieval, and statutory interpretation—and have even achieved passing scores on professional licensing exams, including the U.S. and Brazilian Bar Examinations (Katz et al., 2024; Pires et al., 2025). However, this progress is countered by documented systemic reasoning failures, including hallucinated justifications and non-entailed legal conclusions (Charlotin, 2026; Ariai et al., 2025; Siino et al., 2025). This juxtaposition creates a paradox: models that achieve high benchmark scores often exhibit deep, recurrent reasoning flaws.

Although current evaluations increasingly emphasize process-oriented analyses rather than terminal correctness alone (Guha et al., 2023; Fan et al., 2025), this paradox persists because these approaches remain largely observational, offering limited control over how reasoning unfolds and stabilizes. We contribute to this ongoing shift by adopting *pedagogical scaffolding* (Wood et al., 1976) as an experimental paradigm that progressively injects expert-validated legal guidance into the reasoning process, treating a model's chain-of-thought as an object of controlled analysis rather than a static artifact. We leverage gold-annotated decompositions of Legal QA problems into the canonical FIRAC structure – a widely taught structure that organizes legal reasoning into Facts, Issue, Rules, Application, and Conclusion (White, 2004). Starting from no guidance, we incrementally provide these components in a cumulative manner, such that each step strictly adds legally relevant information without removing prior context. This design, illustrated in Figure 1, induces a controlled monotonic guidance protocol by construction; each model produces a

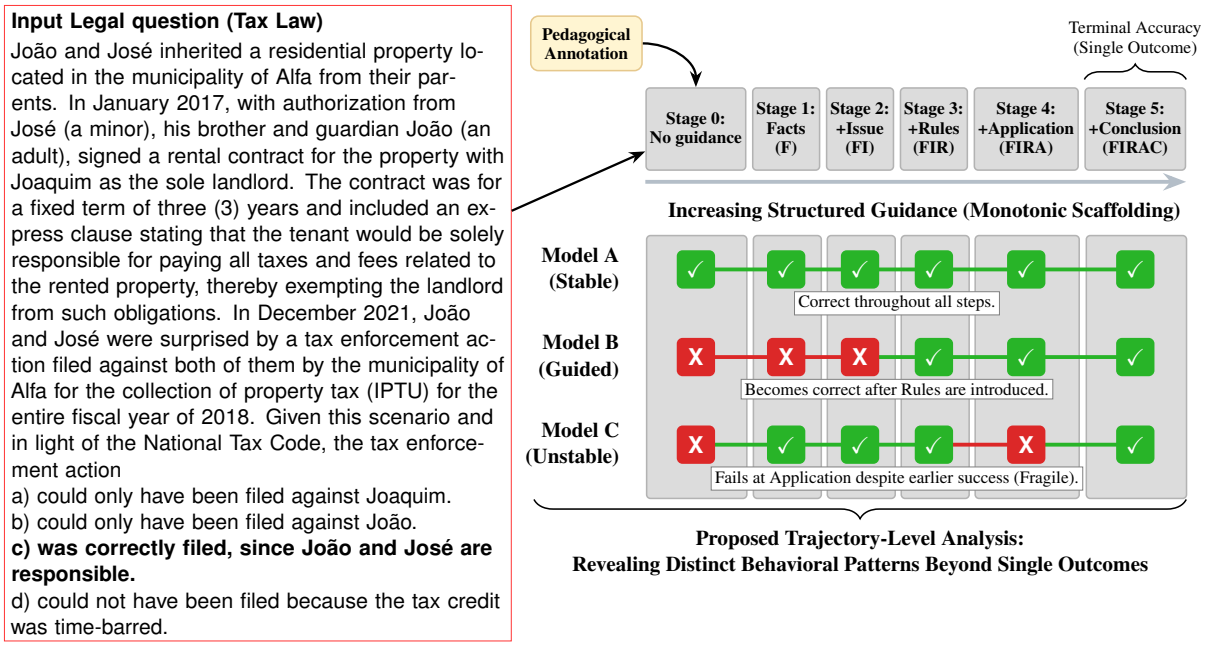

Figure 1: **Overview of the monotonic scaffolding methodology**. Model-specific reasoning trajectories are elicited via progressively legal guidance. Model A remains correct throughout; Model B becomes correct only after legal rules are introduced (FIR); and Model C initially fails, succeeds under intermediate guidance, but exhibits a regression at the application stage (FIRA), illustrating trajectory instability under increasing guidance.

sequence of responses to a legal question under increasing levels of legal guidance. We treat this ordered sequence as a *reasoning trajectory*.

Framing reasoning in this way enables direct intervention on the reasoning process itself, allowing us to analyze not only whether a model ultimately reaches the correct conclusion, but how correctness emerges, stabilizes, or deteriorates as additional legal guidance is introduced. This perspective exposes behavioral dimensions that terminal accuracy alone cannot capture. Specifically, we investigate:

1. **Do different language models operate in qualitatively distinct reasoning regimes?** Can models be characterized according to their responsiveness to the levels of guidance and the stability of their intermediate reasoning beyond differences in aggregate accuracy?

2. **How can trajectory-based evaluation metrics reduce overestimation of model ability caused by guessing?** In particular, does requiring trajectory-level stability provide a stricter notion of correctness under monotonic scaffolding?

3. **What systematic patterns of reasoning stability and failure emerge across legal domains?** In particular, do outcome-level performance and process-level consistency organize differently across domains (e.g., Civil Law, Human Rights Law), revealing domain-dependent latent structures?

Leveraging the monotonic nature of the intervention, we introduce two local transition-based metrics: *Errors-to-Success (E2S)* and *Success-to-Errors (S2E)*, that respectively capture how much guidance models require to reach correctness and how fragile that correctness becomes as guidance increases. We further derive *Stable Accuracy*, a global robustness metric that credits correctness only when a model answers correctly across all scaffolding stages and highlight differences of stability among models.

We instantiate a controlled monotonic scaffolding experiment over 17 language models using 3,123 Brazilian Bar Exam multiple-choice questions spanning 20 legal domains, each paired with gold-standard explanations aligned with the FIRAC legal reasoning structure. Applying our proposed metrics, we find that while frontier models such as `gpt-5.2` exhibit highly stable reasoning trajectories, smaller and open-weight models display substantially higher transition volatility. Moreover, Stable Accuracy reduces the absolute performance of mid-sized models by up to 30 percentage points, indicating that this metric more accurately captures their lack of robustness.

Finally, we apply principal component analysis (PCA) to transition profiles aggregated by domain. This yields a low-dimensional representation revealing that legal domains differ systematically in their reliance on the various guidance levels. These results suggest that reasoning difficulty is non-uniform and shaped by domain-specific structural demands, demonstrating that monotonic scaffolding provides both local behavioral diagnostics and global insights into the latent structure of legal reasoning.

2 Related Work

The intersection of Law and NLP is no longer speculative. As a fundamentally text-heavy domain, law relies on large volumes of statutes, case law, contracts, and regulations, making it a natural target for NLP systems that partially automate core legal tasks (Kang et al., 2025; Schwarcz et al., 2025; Waisberg and Hudek, 2021). At the same time, legal reasoning depends on dense, interrelated texts, long-range dependencies, and strict justificatory standards, making law a particularly demanding testbed that exposes both the strengths and limitations of NLP methods in reasoning-intensive contexts (Nguyen et al., 2025; Nazarenko and Wyner, 2017; Dugac and Altwicker, 2025).

In response, the NLP community has developed a growing ecosystem of Legal Question Answering (QA) benchmarks aimed at assessing whether models can follow legal logic, apply legal rules, and engage in structured argumentation across realistic scenarios (Chlapanis et al., 2025; Abdallah et al., 2023; Fei et al., 2024; Zheng et al., 2025). Resources such as LegalBench (Guha et al., 2023) and LEXam (Fan et al., 2025) provide broad coverage of legal tasks, while more recent efforts incorporate explicit legal structures—such as IRAC-style components—into evaluation, signaling a shift toward process-aware assessment (Yu et al., 2025). Collectively, these works reflect growing recognition that final-answer accuracy alone is insufficient to characterize legal reasoning quality (Jang et al., 2025; Yu et al., 2025).

Despite this progress, most existing approaches treat task decomposition primarily as a descriptive artifact rather than a controlled experimental variable. Legal structures such as IRAC or FIRAC are typically used to annotate outputs or analyze reasoning *post hoc*, after a response has already been generated. As a result, intermediate reasoning steps are observed rather than manipulated, limiting insight into how model behavior changes under increasing guidance. Moreover, high benchmark performance is often driven by models' internalized statutory knowledge rather than a transferable capacity for legal reasoning (Oh et al., 2025). Even process-aware evaluations therefore struggle to distinguish genuine reasoning improvements from brittle or unstable responses.

In contrast, we treat legal task decomposition as an explicit intervention on the reasoning process itself. Drawing on pedagogical scaffolding from educational psychology—where structured support is progressively introduced to reveal and diagnose gaps in understanding (Wood et al., 1976; Van de Pol et al., 2010; Reiser, 2018)—we operationalize legal structure as a controllable source of guidance. This interventionist perspective aligns with emerging dynamic evaluation paradigms in other domains, including interactive program synthesis (Rontogiannis et al., 2025) and reasoning-graph perturbations that probe model robustness by modifying intermediate logic (Zhang et al., 2024; Wang et al., 2024).

Concretely, we adopt FIRAC (Facts, Issue, Rule, Application, Conclusion) (Morgan-Thomas, 2012) as a monotonic scaffold, in which successive components of legally valid reasoning are incrementally introduced as controlled interventions. Guidance is strictly additive: each stage provides additional, expert-validated legal structure without removing prior information. Unlike standard prompting approaches that optimize for end-task performance, this design enables controlled probing of intermediate reasoning states, allowing us to identify the minimal guidance required for correctness and to assess whether that correctness remains stable—or deteriorates—as further normative guidance is introduced (Wang, 2025).

3 Dataset Construction

The Brazilian Bar Examination (OAB) is a professional licensing exam that all law graduates must successfully complete in order to practice law in Brazil. Each edition attracts over 100,000 examinees, with approval rates typically ranging between 20% and 45% (Pires et al., 2025). It is administered three times per year and is composed of two stages: an initial objective phase followed by a written phase in which candidates must demonstrate applied legal reasoning through open-ended

responses. In this study, we restrict our analysis to 43 editions of the objective phase of the exam, each edition comprised of 80 multiple-choice questions spanning a broad range of legal domains, including Constitutional Law, Criminal Law, Administrative Law, Civil Law, Labor Law, Tax Law, Legal Philosophy, and Professional Ethics. All multiple-choice questions are publicly available[1]. Each question is presented as a legal case vignette that requires candidates to identify the legally correct outcome, offered as four-answer options (A–D), as illustrated in Figure 1.

In addition to the questions and their correct alternatives, our analysis requires access to expert-based structured legal guidance that explicates the underlying reasoning process. To this end, we identified a specialized website[2] dedicated to preparing candidates for the Brazilian Bar Exam, which provides detailed pedagogical annotations authored by law professors. These annotations not only justify why the correct alternative is legally valid, but also explain why each incorrect option fails, citing statutory provisions, doctrinal principles, or jurisprudential interpretations and how they should be applied to answer the question.

We then transformed these expert-written explanations into the FIRAC legal structure (Facts, Issue, Rules, Application, Conclusion) (dos Santos et al., 2025). Specifically, for each question, we provided the original question text together with the annotated answer key to a large language model (`gemini 2.5 flash` with default decoding parameters), which was instructed solely to organize the existing expert content into the FIRAC framework. The model did not introduce new legal content or interpretations; it merely segmented and labeled the expert-provided material into the corresponding FIRAC components. This process yields, for every question, a canonical FIRAC-structured solution that preserves the substantive legal reasoning articulated by human experts while making intermediate reasoning steps explicit and standardized. These FIRAC annotations act as ground-truth for chain-of-thought (CoT) reasoning and form the basis for defining scaffolding levels, enabling controlled analysis of model behavior under progressively richer legal guidance.

As a validation step, we performed automatic sanity checks verifying expected lexical markers (e.g., person names in Facts, question marks in Issue, law references in Rule, "therefore" in Application, "conclude" in Conclusion) and markers cluster as expected. Also, for a sample of 100 questions, one author manually verified whether laws mentioned in the annotated answers appeared in the Rule segment of the extracted FIRAC structure and found 0.14% errors. The prompt used to generate FIRAC stages, along with examples of annotated questions and additional details, are provided in Appendices B and G. Our final dataset comprises approximately 3,000 questions written in Brazilian Portuguese, drawn from 2010 to 2025 editions of the exam.

4 Methods

Each legal question is evaluated under the Cartesian product of (i) a set of language models M and (ii) a set of $L{+}1$ scaffolding levels. Formally, for each question q, we observe a sequence of binary outcomes:

$$\left\{ y_q^{(m,\ell)} \;\middle|\; m \in \{1,\dots,M\},\ \ell \in \{0,\dots,L\} \right\}, \tag{1}$$

where $y_q^{(m,\ell)} \in \{0,1\}$ indicates whether model m answers a multiple-choice legal question q correctly when prompted under scaffolding level ℓ.

Scaffolding levels are aligned with the canonical FIRAC structure and are defined cumulatively. Specifically, $\ell = 0$ corresponds to the *no-guidance* condition, in which the model receives only the original question text. Level $\ell = 1$ provides the case *Facts* (F); $\ell = 2$ provides *Facts* + *Issue* (F+I); $\ell = 3$ provides *Facts* + *Issue* + *Rules* (`F+I+R`); and higher levels continue analogously until the full FIRAC structure is revealed. In our experiments $L = 6$; to help disentangle memory from reasoning, we add an extra level before `R` which we name `L`, which gives as guidance just the law identifier (e.g. Federal Law 9.394/1996), instead of its content, which is only provided in the `R` level. Because each scaffolding level strictly adds expert-validated, solution-relevant information without removing or altering prior content, the intervention is *strictly monotonic*, and thus, for a model m, the sequence $\{y_q^{(m,\ell)}\}_{\ell=0}^{L}$ defines a controlled reasoning trajectory in which guidance can only increase.

This representation allows us to apply a range of standard statistical analyses to trajectories of 0s and 1s, including aggregation across levels and the study of transitions between states. Aggregate accuracy summarizes overall performance at each level,

[1] https://oab.fgv.br/.

[2] www.oabnamedida.com.br.

while transition patterns capture how correctness emerges, persists, or degrades as additional guidance is introduced. We introduce two transition-based diagnostics defined over individual model–question trajectories.

Errors-to-Success measures the amount of guidance required for a model to reach correctness. For model m on question q,

$$\mathrm{E2S}(q, m) = \min\{\ell : y_q^{(m,\ell)} = 1\}. \quad (2)$$

Lower E2S values indicate that less guidance is needed for a model to produce a correct answer; specifically, if $E2S = 0$ the model got the correct answer without any guidance.

Success-to-Errors (S2E), on the other hand, quantifies the fragility of correctness under additional guidance. Let $\ell_{\text{first}} = \min\{\ell : y_q^{(m,\ell)} = 1\}$ denote the first success level. Then,

$$\mathrm{S2E}(q, m) = \min\{\ell > \ell_{\text{first}} : y_q^{(m,\ell)} = 0\} - \ell_{\text{first}}. \quad (3)$$

Lower S2E values indicate greater stability: once correctness is achieved, it is not followed by subsequent failures. In Figure 1, the trajectories are 111111, 000111, and 011101 and the corresponding E2S values are 0, 3, and 1, while the S2E values are 0, 0, and 3, respectively. The first two trajectories exhibit stable correctness once achieved (no degradations), whereas the third shows a failure post-success, indicating that correctness is fragile and not preserved under additional guidance.

Stable Accuracy. Stable Accuracy assigns a question as correct for a given model if and only if correctness is achieved immediately and maintained throughout the entire scaffolding trajectory. Formally, a trajectory is counted as correct when $\mathrm{E2S}(q, m) = 0$ and $\mathrm{S2E}(q, m) = 0$.

Together, E2S, S2E and Stable Accuracy transform outcome sequences into interpretable statistics that quantify both the need for guidance and the robustness of correctness under monotonic interventions. Conceptually, these diagnostics are related to cumulative scaling approaches in psychometrics, such as the Guttman scale (Guttman, 1944), which characterize deviations from response patterns across ordered conditions.

Trajectory Analysis with PCA. Representing model outcomes as trajectories, as defined in Eq. 1, enables the application of a broad class of multivariate analysis techniques, including clustering, anomaly detection, and dimensionality reduction. In this analysis, we construct a trajectory representation per question, where each dimension corresponds to performance at a given scaffolding level aggregated across models. We then apply Principal Component Analysis (PCA) as an exploratory tool over this question-level trajectory space. Our goal is to examine whether questions from different legal domains—such as Tax Law and Civil Law—systematically occupy distinct regions of the resulting latent space, revealing domain-specific patterns in how models collectively respond to increasing legal guidance.

5 Results

We evaluate a range of OpenAI models—`gpt-4o-mini` and `gpt-5.2`, `gpt-5-mini`, and `gpt-5-nano` with different `reasoning.effort`[3] settings. We also evaluate four instruction-tuned Gemma-3 models, in their e4b, 4b, 12b and 27b versions. In addition, we experiment with Brazilian-focused models: (i) `Sabia 3` and 3.1, and its smallest counterpart, `Sabiazinho-3` (Abonizio et al., 2025); and (ii) `Jurema-7B`[4], a fine-tuned variant of `Qwen-7B-Instruct` specialized in Brazilian law. Finally, we include `Qwen2.5-7B-Instruct` and `Qwen2.5-14B-Instruct` models for comparison.

We organize our discussion around three directions that directly correspond to the three research questions outlined in Section 1: monotonic scaffolding (1) qualitatively differentiates models, (2) yields more robust accuracy metrics with reduced overestimation of model abilities, and (3) structures legal domains according to the degree of normative knowledge they require.

5.1 Model Differentiation under Monotonic Scaffolding

As a first-order summary of model behavior, we aggregate terminal outcomes to obtain conventional accuracy metrics stratified by both scaffolding level and model in Figure 2. Across the board, performance improves monotonically as additional gold-standard legal structure is introduced, but the magnitude and smoothness of these gains vary substantially by model family and scale. Frontier models (`GPT-5.2` variants) already achieve high accuracy in the unstructured setting and exhibit relatively shallow but stable gains, converging near ceiling

[3] https://platform.openai.com/docs/guides/latest-model.
[4] https://huggingface.co/Jurema-br/Jurema-7B.

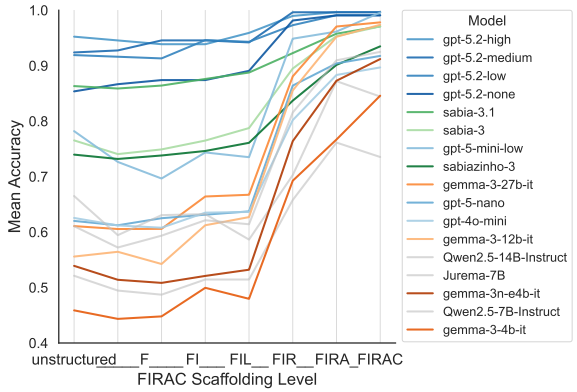

Figure 2: Mean accuracy increases consistently with additional FIRAC components, demonstrating a systematic average benefit of structured scaffolding. Exact numbers are documented in Appendix D.

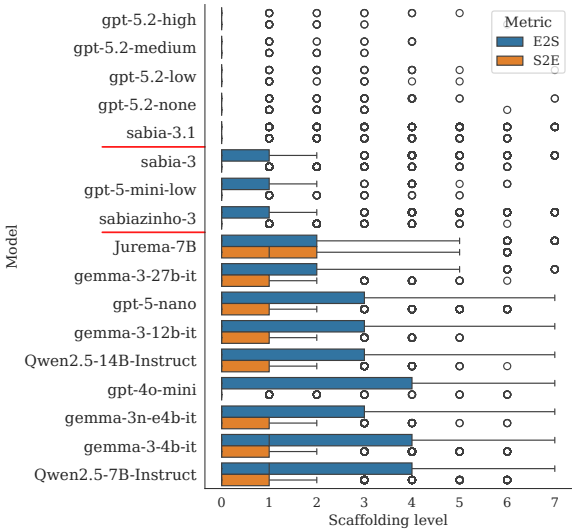

Figure 3: Distributions of Errors-to-Success (E2S) and Success-to-Errors (S2E) across models. The joint view reveals distinct performance regimes, separating models that require substantial scaffolding to reach correctness from those that exhibit early success and higher reasoning stability.

performance once rules and application-level guidance (*FIR–FIRA*) are provided. In contrast, mid-sized and smaller models show pronounced sensitivity to scaffolding: their trajectories display modest or noisy improvements on early stages (*F*, *FI*, *FIL*), followed by sharp accuracy jumps once explicit legal rules and applications are introduced. This pattern suggests that, for these models, correctness is often unlocked only after substantial normative guidance. Notably, no model exhibits systematic degradation under added guidance, supporting the monotonicity assumption of the intervention.

The aggregated trajectories displayed in Figure 2 also expose specific model fragilities hidden in terminal accuracy alone. We can see that, in the final FIRAC level—where the model receives full information, including the expected conclusion of the legal case, guided residual errors are less plausibly attributable to a lack of domain knowledge. Instead, they betray a fundamental fragility in instruction following. For instance, despite receiving exhaustive scaffolding, `gemma-3-4b-it` reaches only 82% accuracy, making it clear that the model has a strong fragility in performing a simple mapping between “Conclusions" and the answer option. This is a failure mode that remains largely obscured when evaluating its final accuracy alone (45%).

E2S and S2E. In Figure 3, models are compared according to the distribution of E2S and S2E computed over all questions, revealing a clear stratification into three regimes. Frontier model variants based on `gpt-5.2` form the top regime, with both E2S and S2E sharply concentrated around zero, indicating that these models typically reach correctness immediately and maintain it robustly under additional scaffolding. A second regime comprises `gpt-5-mini` and the `sabia-3` family (including `sabiazinho-3`), which remain highly stable but exhibit slightly higher error rates: their E2S values are concentrated in the 0–1 range, reflecting occasional initial failures before convergence, while S2E remains near zero. In contrast, the remaining models—including Gemma variants and `gpt-5-nano`—occupy a lower regime characterized by both higher E2S, often spanning the full 0–4 range, and noticeable instability, with S2E values concentrated in the 0–1 range. This indicates that these models not only require substantially more guidance to reach correctness, but also fail to consistently preserve correctness once achieved. Taken together, E2S and S2E expose qualitatively distinct reasoning regimes that are not apparent under standard accuracy-based evaluation.

To verify whether the observed patterns would vanish under repeated sampling, we evaluated a majority-vote baseline over three independent runs with shuffled answer option orders. The trajectory-level metrics remain stable under aggregation, indicating that the dynamics are not artifacts of stochastic decoding. As expected, E2S increases slightly under majority voting (e.g., 2.05 vs.1̃.92 for `Qwen2.5-7B`), since aggregation removes occasional early successes driven by sampling vari-

ance. Conversely, S2E decreases modestly (e.g., 0.91 vs.1̃.04 for `Jurema-7B`), as majority voting suppresses sporadic regressions after an initially correct answer. Crucially, trajectory regressions persist even after aggregation, suggesting that these effects reflect a structural sensitivity of model reasoning to additional scaffolding rather than noise.

Diagnosing the effects of fine-tuning. Compared to its base model, `Qwen2.5-7B-Instruct`, `Jurema-7B` shows substantially lower Errors-to-Success (E2S) values (0–1 vs. 0–4), indicating that fine-tuning effectively injects legal knowledge and allows the model to reach correct answers with less scaffolding. Its E2S performance is even stronger than that of the much larger `Qwen2.5-14B-Instruct` (0–3), suggesting that domain specialization can outweigh differences in model scale. However, this gain comes with a trade-off. The Success-to-Errors (S2E) metric reveals that `Jurema-7B` is less robust: its S2E values extend up to 0–2, compared to 0–1 for `Qwen2.5-7B-Instruct`, indicating more frequent regressions after initially correct answers. Such results help elucidate concerns that fine-tuning processes may inadvertently degrade previously acquired capabilities (Luo et al., 2023). Appendix F complements these results with qualitative examples where `Jurema-7B` fails while the other models do not, and we found that misrepresented norms are by far the biggest offender.

We view this analysis as revealing regime-level differences across model families when evaluated under monotonic scaffolding. Instead of reflecting on whether an improvement of a certain number of percentage points is relevant for the sake of an specific application, Figure 3 enables more conscious cost–benefit decisions in model selection. Models that achieve early and stable correctness are better suited to high-stakes legal tasks requiring robustness, while models that rely on deeper scaffolding may be sufficient and more economical for tasks where additional structure can be provided and occasional instability is acceptable.

5.2 Stable Accuracy beyond Overestimation

A central concern in psychometrics is disentangling true ability from guessing (Żółtak and Golonka, 2015). In Legal QA, this concern is amplified: a correct final answer may result from partial pattern matching or opportunistic elimination rather than stable legal reasoning. To address this, we introduce Stable Accuracy, a stricter criterion that

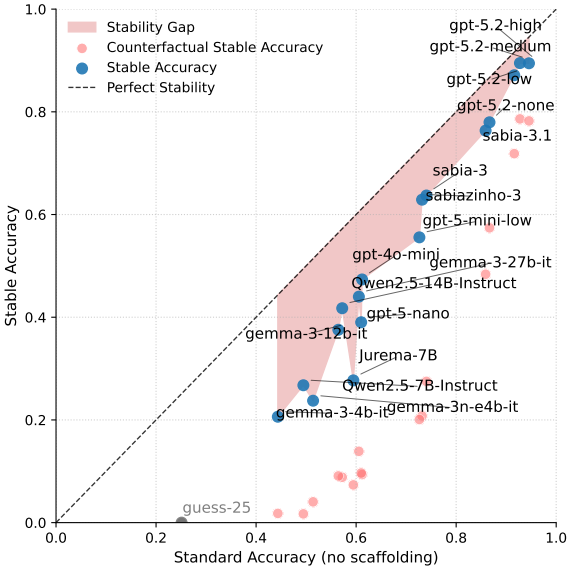

Figure 4: **Standard vs. Stable Accuracy**. The distance below the $y = x$ diagonal represents the instability gap, quantifying the extent to which standard evaluation overestimates legal reasoning by rewarding brittle or opportunistic successes.

credits a response as correct only if correctness is maintained across the entire monotonic scaffolding trajectory. Consider a model that purely guesses, `guess-25`. Such a model has an expected *Standard Accuracy* of 0.25 on four-option questions. However, its expected *Stable Accuracy* decays exponentially with the number of scaffolding levels ℓ, namely, the probability of remaining correct throughout the trajectory is 0.25^{ℓ}. For even modest ℓ, this value is effectively zero.

Figure 4 plots Stable Accuracy (y-axis) against standard terminal accuracy without scaffolding (x-axis). The diagonal $y = x$ represents a frontier where terminal correctness is fully supported by trajectory-level stability. Models near this frontier exhibit little degradation when moving from standard to stable evaluation, indicating that their performance reflects coherent and persistent reasoning rather than isolated successes.

Two systematic patterns emerge. First, standard accuracy consistently overestimates true legal ability. Nearly all models fall below the diagonal, showing that a nontrivial portion of their apparent performance fails to survive stability constraints. This gap quantifies how much standard evaluation inflates perceived competence by crediting fragile or non-reproducible successes.

Second, instability is universal but highly unequal. All models exhibit some degree of degra-

dation, but weaker models fall disproportionately farther from the $y = x$ frontier. Their larger stability gaps indicate a heavier reliance on opportunistic correctness—answers that do not generalize even under minimal increases in legal structure. In contrast, stronger models degrade more gracefully, suggesting qualitatively different reasoning regimes rather than mere differences in average accuracy.

The plot also reveals a temporal trend: more recent models are not only more accurate, but more robust. This indicates that recent gains are accompanied by improved consistency under structured guidance, rather than being driven solely by increased surface-level performance.

We further validate this interpretation through a trajectory-shuffling control experiment. By randomly permuting correctness trajectories within the same model and FIRAC level—thereby destroying temporal coherence while preserving marginal accuracy — we observe a sharp collapse in Stable Accuracy (orange points). This confirms that Stable Accuracy is sensitive to structured, coherent reasoning dynamics rather than aggregate correctness alone: once trajectory structure is removed, performance resembles near-guessing behavior.

Taken together, these results establish Stable Accuracy as a psychometrically stronger evaluation metric than Standard Accuracy. It exposes systematic overestimation of legal ability, reveals disproportionate instability in weaker models, and demonstrates that recent model improvements reflect genuine gains in robustness rather than superficial accuracy alone.

5.3 Normative Demand across Legal Domains

In Sections 5.1 and 5.2 we show model-level patterns. Here, we focus on the questions. Do different legal domains require different levels of guidance? To address this, we conduct a principal component analysis aimed at identifying the dimensions along which questions vary in their sensitivity to scaffolding, and at examining whether legal domains organize meaningfully along these dimensions.

We analyze the multivariate structure of question trajectories using Principal Component Analysis (PCA), as illustrated in Figure 5. The first principal component (PC1) explains 69% of the total variance, while the second component (PC2) explains an additional 14%, together accounting for over 80% of the variability in the trajectory space. PCA Loadings are displayed in Appendix E; all FIRAC-level loadings on PC1 are positive, and PC1

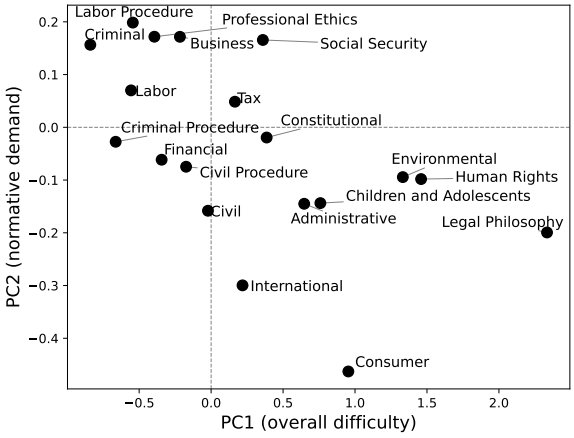

Figure 5: Scatter plots of the first two principal components (PC1 and PC2), showing centroids of questions aggregated by legal subfield.

correlates almost perfectly with standard accuracy ($r = 0.99$). This indicates that PC1 primarily captures a notion of global question difficulty: questions with higher PC1 scores are easier across all scaffolding levels, while lower scores correspond to uniformly difficult questions.

In contrast, PC2 exhibits a markedly different structure. It is positively weighted by the `FIR`, `FIRA`, and `FIRAC` dimensions, and negatively weighted by the remaining levels. This pattern suggests that PC2 captures sensitivity to normative structure and rule application. Questions with positive PC2 scores benefit disproportionately from the introduction of explicit legal rules and their application. As shown in Figure 5, this dimension separates domains characterized by dense, technical, and highly structured bodies of law—such as Criminal Law, Social Security Law, and Labor Procedure—from domains such as Legal Philosophy and International Law, which are guided more by broad principles and general doctrines rather than by detailed and intricately coupled norms. In Appendix G we show examples of questions on each extreme of the PC1-PC2 space. Beyond evaluation, this decomposition has direct pedagogical implications. In instructional settings with law students, PC1-like effects correspond to overall difficulty and can inform sequencing and pacing of assessments, while PC2 captures how legal knowledge is mobilized—whether mastery depends on the application of detailed norms or on principled, conceptual reasoning. As a result, trajectory-based PCA analyses, grounded in Figure 5, provide a principled way to design targeted scaffolding, diagnose domain-

specific learning challenges, and align legal education more closely with the cognitive demands of different areas of law.

6 Conclusions

Large language models exhibit a paradox in legal reasoning: they achieve strong performance on bar-exam benchmarks while still producing hallucinated justifications and unstable reasoning. We contribute to clarifying this tension by proposing an evaluation protocol based on monotonic scaffolding, instantiated through the FIRAC framework (Facts, Issue, Rules, Application, Conclusion) and treated as a controlled experimental intervention rather than a prompting heuristic. By examining model behavior across ordered FIRAC levels, we introduce transition-based diagnostics (E2S, S2E) and Stable Accuracy, which reveal reasoning instability and reduce overestimation due to guessing. Our results show that models operate in qualitatively distinct reasoning regimes and that legal domains organize according to their normative demands—patterns that remain opaque under outcome-only evaluation.

More broadly, because FIRAC serves here as one instance of a staged reasoning schema, our methodology generalizes beyond law to other domains with structured reasoning, offering a principled foundation for process-aware evaluation focused on stability rather than isolated correctness.

7 Limitations

This study has limitations that point to directions for future work. First, our empirical analysis is conducted on a single legal jurisdiction and a single language. While this may limit immediate generalizability, we partially mitigate this constraint by focusing on Brazilian law, a civil-law system widely recognized for its doctrinal complexity, procedural formalism, and dense statutory structure. In addition, our study contributes linguistic diversity to the literature by examining a non-English legal setting, which remains underrepresented in empirical evaluations of large language models.

Second, our evaluation design incurs higher computational and monetary costs than standard accuracy-based benchmarks. Because each question is evaluated under L monotonic scaffolding levels for every model, the total number of executions scales linearly with both the number of models and the number of scaffolding conditions. This cost is an inherent trade-off of trajectory-level analysis: observing reasoning dynamics requires repeated, structured probing rather than a single terminal response.

Third, some of the empirical patterns reported in this study may be contingent on specific combinations of model architecture, prompt formulation, and scaffolding design. Although our scaffolding levels are constructed to be monotonic and content-preserving by design, different prompt realizations or alternative decompositions of the reasoning stages could induce different transition dynamics, particularly for smaller or instruction-sensitive models. As a result, the absolute values of metrics such as E2S and S2E should be interpreted as conditional on the evaluated model–prompt–scaffolding configuration, rather than as intrinsic properties of a model in isolation.

A related question is whether the task may in fact become more difficult for models as additional guidance is provided, since each scaffolding step introduces new semantic and normative constraints that the model must integrate consistently, rather than merely supplying redundant hints that simplify the problem. In our dataset, the unguided questions are short, averaging approximately 300 tokens, while the fully scaffolded prompts with complete FIRAC annotations reach roughly 1,000 tokens, around 3% of the support context window of Sabia models, for example (Abonizio et al., 2025). This range remains within a small to moderate context regime for modern language models and is well below their maximum context windows, suggesting that the observed effects are unlikely to be driven by long-context phenomena.

Finally, our framework depends on the availability of gold-standard step-level annotations that define the scaffolding stages. In our case, these stages are derived from the FIRAC legal reasoning schema and require expert curation to ensure consistency and validity. Producing such annotations is labor-intensive and may be challenging to replicate at scale or in domains without well-established reasoning frameworks. Although we argue that many structured domains admit analogous schemas, the effort required to instantiate high-quality step-level supervision remains a practical limitation of the approach.

8 Ethical Considerations

While the proposed methodology is diagnostic rather than deployable, we raise two main ethical considerations and risks.

Use of Legal Data. The legal questions used in this study are drawn from publicly available Brazilian Bar Examination materials. However, the expert-annotated solutions that provide structured legal reasoning are not publicly released as part of the original exam materials. These annotations originate from third-party educational resources and reflect pedagogical explanations authored by legal experts. We use these materials strictly for research and evaluation purposes, without redistributing the annotated content. The dataset contains no personal identifiers, sensitive personal data, or private communications. We emphasize that both the questions and the associated annotations are intended for educational use and should not be construed as authoritative legal advice.

Bias and Jurisdictional Scope. The study is limited to Brazilian law, a single legal system with its own doctrinal and procedural characteristics. Conclusions about model robustness and reasoning stability may not transfer directly to other jurisdictions or legal traditions. There is a risk of implicit bias if results are generalized beyond this context. We therefore frame our claims narrowly and emphasize that the methodology—rather than the empirical rankings of models—is the primary contribution.

9 Acknowledgements

Supported by CNPq, CAPES, FAPEMIG, IAIA-INCT on AI and INCT-TILDIAR.

References

Abdelrahman Abdallah, Bhawna Piryani, and Adam Jatowt. 2023. Exploring the state of the art in legal qa systems. *Journal of Big Data*, 10(1):127.

Hugo Abonizio, Thales Sales Almeida, Thiago Laitz, Roseval Malaquias Junior, Giovana Kerche Bonás, Rodrigo Nogueira, and Ramon Pires. 2025. Sabiá-3 technical report. *Preprint*, arXiv:2410.12049.

Farid Ariai, Joel Mackenzie, and Gianluca Demartini. 2025. Natural language processing for the legal domain: A survey of tasks, datasets, models, and challenges. *ACM Computing Surveys*, 58(6):1–37.

Damien Charlotin. 2026. Ai hallucination cases database. https://www.damiencharlotin.com/hallucinations/. Accessed: 2026-01-05.

Odysseas S. Chlapanis, Dimitrios Galanis, Nikolaos Aletras, and Ion Androutsopoulos. 2025. Greekbarbench: A challenging benchmark for free-text legal reasoning and citations. *Preprint*, arXiv:2505.17267.

Farhad Dehghani, Reza Dehghani, Yasaman Naderzadeh Ardebili, and 1 others. 2025. Large language models in legal systems: A survey. *Humanities and Social Sciences Communications*, 12:1977.

Janderson Glauber Mendes dos Santos, Pedro Calais, Wagner Meira Jr, Ricardo Shen, Aziz Tuffi Saliba, Anisio Mendes Lacerda, Ana Luísa Vaz Barbosa Araújo, Fernanda Alves de Carvalho, Gabriel Rolla Ferreira, and Victor Augusto Hon Fonseca. 2025. Evaluation of legal reasoning in language models based on the IRAC structure. In *Proc. of the XVII Brazilian Congress of Computational Intelligence (CBIC 2025)*, pages 1–8, Belo Horizonte, MG. SBIC.

Gaspar Dugac and Tilmann Altwicker. 2025. Classifying legal interpretations using large language models. *Artificial Intelligence and Law*, pages 1–19.

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, Jidong Ge, and Vincent Ng. 2024. LawBench: Benchmarking legal knowledge of large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7933–7962. Association for Computational Linguistics.

Neel Guha, Julian Nyarko, Daniel Ho, Christopher Ré, Adam Chilton, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel Rockmore, Diego Zambrano, 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.

Louis Guttman. 1944. A basis for scaling qualitative data. *American sociological review*, 9(2):139–150.

Yehoon Jang, Chaewon Lee, Hyun-seok Min, and Sungchul Choi. 2025. Pilot-bench: A benchmark for legal reasoning in the patent domain with irac-aligned classification tasks. In *Proceedings of the Natural Legal Language Processing Workshop 2025*, pages 240–280.

Xiaoxi Kang, Lizhen Qu, Lay-Ki Soon, Zhuang Li, and Adnan Trakic. 2025. Automating irac analysis in malaysian contract law using a semi-structured knowledge base. *Artificial Intelligence and Law*, pages 1–44.

Daniel Martin Katz, Michael James Bommarito, Shang Gao, and Pablo Arredondo. 2024. Gpt-4 passes the bar exam. *Philosophical Transactions of the Royal Society A*, 382(2270):20230254.

Yun Luo, Zhen Yang, Fandong Meng, Yafu Li, Jie Zhou, and Yue Zhang. 2023. An empirical study of catastrophic forgetting in large language models during continual fine-tuning. *IEEE Transactions on Audio, Speech and Language Processing*, 33:3776–3786.

Maxine Morgan-Thomas. 2012. The legal studies case brief assignment: Developing the reading comprehension bridge to critical thinking. *International Journal of Business and Social Science*, 3(23).

Adeline Nazarenko and Adam Wyner. 2017. Legal nlp introduction. *Traitement automatique des langues*, 58(2):7–19.

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.

Hongseok Oh, Wonseok Hwang, and Kyoung-Woon On. 2025. Korean canonical legal benchmark: Toward knowledge-independent evaluation of llms' legal reasoning capabilities. *arXiv preprint arXiv:2512.24572*.

Ramon Pires, Roseval Malaquias Junior, and Rodrigo Nogueira. 2025. Automatic legal writing evaluation of llms. In *Proceedings of the International Conference on Artificial Intelligence and Law (ICAIL)*.

Brian J Reiser. 2018. Scaffolding complex learning: The mechanisms of structuring and problematizing student work. In *Scaffolding*, pages 273–304. Psychology Press.

Dimitrios Rontogiannis, Maxime Peyrard, Nicolas Baldwin, Martin Josifoski, Robert West, and Dimitrios Gunopulos. 2025. Interactive evaluation of large language models for multi-requirement software engineering tasks. *arXiv preprint arXiv:2508.18905*.

Daniel Schwarcz, Sam Manning, Patrick Barry, David R Cleveland, JJ Prescott, and Beverly Rich. 2025. Ai-powered lawyering: Ai reasoning models, retrieval augmented generation, and the future of legal practice.

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.

Janneke Van de Pol, Monique Volman, and Jos Beishuizen. 2010. Scaffolding in teacher–student interaction: A decade of research. *Educational psychology review*, 22(3):271–296.

N Waisberg and S Hudek. 2021. Ai for lawyers: How artificial intelligence is adding value. *Amplifying Expertise, and Transforming Careers*.

Charles L Wang. 2025. Mathbode: Measuring the stability of llm reasoning using frequency response. *arXiv preprint arXiv:2509.23143*.

Siyuan Wang, Zhongyu Wei, Yejin Choi, and Xiang Ren. 2024. Can llms reason with rules? logic scaffolding for stress-testing and improving llms. *arXiv preprint arXiv:2402.11442*.

Nancy J White. 2004. Using law class to teach problem-solving and writing skills.

David Wood, Jerome S Bruner, and Gail Ross. 1976. The role of tutoring in problem solving. *Journal of child psychology and psychiatry*, 17(2):89–100.

Yang Wu, Chenghao Wang, Ece Gumusel, and Xiaozhong Liu. 2024. Knowledge-infused legal wisdom: Navigating LLM consultation through the lens of diagnostics and positive-unlabeled reinforcement learning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15542–15555, Bangkok, Thailand. Association for Computational Linguistics.

Wenhan Yu, Xinbo Lin, Lanxin Ni, Jinhua Cheng, and Lei Sha. 2025. Benchmarking multi-step legal reasoning and analyzing chain-of-thought effects in large language models. *arXiv preprint arXiv:2511.07979*.

Zhehao Zhang, Jiaao Chen, and Diyi Yang. 2024. Darg: Dynamic evaluation of large language models via adaptive reasoning graph. *Advances in Neural Information Processing Systems*, 37:135904–135942.

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.

Tomasz Żółtak and Grzegorz Golonka. 2015. Does guessing matter? differences between ability estimates from 2pl and 3pl irt models in case of guessing.

A Experimental Setup

We ran the open-weight models obtained from HuggingFace on Google Colab, incurring a total cost of approximately USD 100. The OpenAI models were accessed via APIs, with a total cost of approximately USD 250. We used the Batch API (https://platform.openai.com/docs/api-reference/batch/create), which substantially reduces inference costs.

All models were evaluated using their default parameters. The only parameter we modified was `reasoning.effort` for the `GPT-5` family, which can be set to `none`, `low`, `medium`, or `high`. As described in https://platform.openai.com/docs/guides/latest-model , *the* `reasoning.effort` *parameter controls how many reasoning tokens the model generates before producing a response*. We varied this parameter to examine its effect on our proposed metrics.

All models are licensed for research purposes; see for example https://huggingface.co/Jurema-br/Jurema-7B.

B Prompts

In this section we show the prompts we have used to produce the results of the paper. All the prompts were originally written in Portuguese, but here we present them in English.

To generate the Legal QA dataset annotated with FIRAC stages, the following prompt was submitted against `gemini 2.5`. Each question is submitted along with its annotated solution produced by legal experts; the task of the prompt is to organize the information into the FIRAC stages. We then have questions paired with expert-annotated solutions.

prompt: Extract FIRAC annotated structure

From the OAB question and the annotated solution below, extract the FIRAC structure (Facts, Issue, Rule, Application, Conclusion) from the legal problem. Restrict yourself solely to what is mentioned in the commented solution, but do not mention the commented solution directly. The output must be in JSON format with the keys "Facts", "Issue", "Rule", "Application", and "Conclusion".
Facts: Detailed identification of the relevant events that gave rise to the legal dispute. Generate a list of facts based on what is mentioned in the question statement. It must be a list ([]).

Issue: Precise formulation of the central legal question to be analyzed and resolved.

Rule: Selection of the legal rules applicable to the problem.

Application: Confrontation of the legal rules with the facts to justify the solution. Do not explicitly mention any answer choice from the question.

Conclusion: Definition of the proposed legal solution based on the analysis of all the previous items. Do not explicitly mention any answer choice from the question.

{{question}}
{{annotated solution}}

The following prompt is used to instruct the model to solve a Brazilian multiple-choice legal question using the FIRAC structure (Facts, Issue, Rule, Application, Conclusion). The model must consider only Brazilian legislation and must explicitly structure its reasoning according to these five components.

The question text is provided verbatim. Some FIRAC fields may already be filled; these fields must not be modified and must be used as given when completing the solution. Any remaining FIRAC fields must be completed in a concise and direct manner.

The response must be returned exclusively in

the specified JSON format, including the "resposta" field, which contains the correct alternative (A/B/C/D). No additional text, explanations, or formatting outside the JSON object are permitted.

prompt: solve MQA question with FIRAC

Solve the legal question below considering Brazilian legislation. Use the FIRAC legal reasoning framework (Facts, Issues, Rules, Application, Conclusion).
Question:

```
{{question}}
```

Respond exclusively in the following JSON format, explaining each FIRAC step concisely and directly. The F/I/R/A/C fields that are already filled in must not be altered, and you must use them as hints to solve the question.

```
{
  "F": "{{F}}",
  "I": "{{I}}",
  "R": "{{R}}",
  "A": "{{A}}",
  "C": "{{C}}",
  "answer": "A/B/C/D"
}
```

In addition to completing any unfilled F/I/R/A/C fields, also fill in the "resposta" field of the JSON with the correct alternative (A/B/C/D). Return only the completed JSON.

JSON:

C Position Bias Check

As a sanity check, we analyze the distribution of selected answer options (A–D) for each model, normalized by model. Across all evaluated models, we observe no systematic preference for specific option positions, indicating the absence of meaningful position bias. This result suggests that subsequent analyses are unlikely to be confounded by superficial answer-order effects.

D Detailed Metrics

We present a detailed breakdown of model performance using complementary, fine-grained metrics. Figure 7 shows that mean accuracy increases monotonically as additional FIRAC components are provided, indicating a systematic average benefit from structured legal scaffolding. Table 1 disaggregates this effect by legal domain and scaffolding stage, revealing substantial heterogeneity across domains and highlighting where gains emerge most strongly along the FIRAC trajectory. Finally, Table 2 reports aggregate accuracy alongside the transition-based diagnostics Errors-to-Success (E2S) and Success-to-Errors (S2E), with confidence intervals, enabling a nuanced comparison of not only how accurate models are, but also how efficiently they reach correctness and how stable that correctness remains under additional structure.

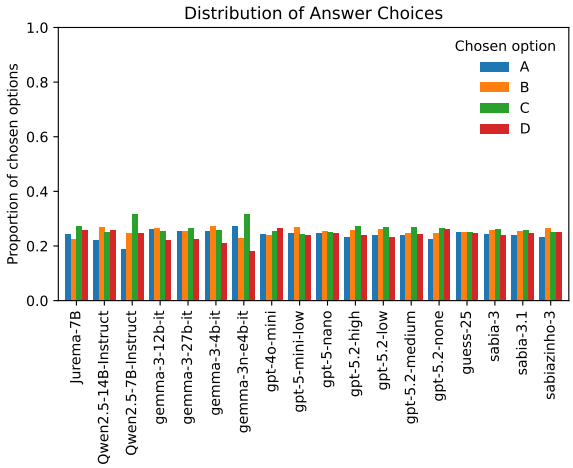

Figure 6: Distribution of selected answer options (A–D) across models, normalized per model. The absence of systematic skew toward specific option positions suggests no meaningful position bias in model answer selection.

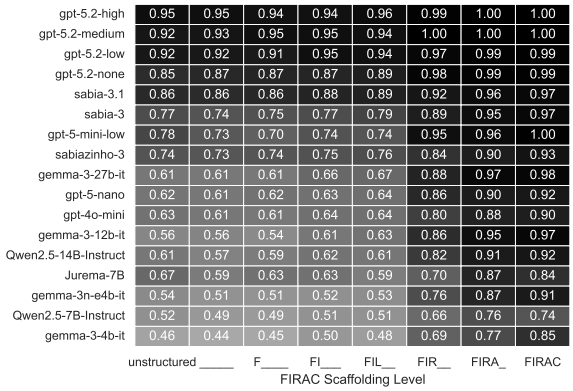

Figure 7: Mean accuracy increases consistently with each additional FIRAC component across all models, demonstrating a systematic average benefit of structured scaffolding.

Legal Domain	_____	F____	FI___	FIL__	FIR__	FIRA_	FIRAC
Legal Philosophy	0.87	0.87	0.89	0.88	0.90	0.92	0.91
Human Rights Law	0.76	0.77	0.79	0.79	0.87	0.89	0.90
Environmental Law	0.77	0.77	0.77	0.77	0.86	0.89	0.88
Consumer Law	0.73	0.75	0.74	0.72	0.79	0.86	0.84
Administrative Law	0.68	0.69	0.71	0.71	0.81	0.87	0.86
Children and Adolescents Law	0.68	0.68	0.70	0.72	0.82	0.86	0.87
Constitutional Law	0.65	0.65	0.67	0.67	0.80	0.85	0.87
Tax Law	0.60	0.62	0.63	0.64	0.78	0.86	0.85
Social Security Law	0.61	0.64	0.61	0.61	0.76	0.83	0.89
Civil Law	0.60	0.61	0.65	0.65	0.75	0.84	0.84
International Law	0.62	0.61	0.63	0.65	0.75	0.82	0.81
Business Law	0.58	0.58	0.61	0.61	0.78	0.83	0.87
Civil Procedure Law	0.59	0.59	0.61	0.60	0.78	0.82	0.85
Financial Law	0.61	0.59	0.61	0.60	0.75	0.84	0.84
Labor Procedure Law	0.55	0.55	0.57	0.57	0.77	0.83	0.86
Professional Ethics	0.53	0.54	0.57	0.57	0.78	0.84	0.85
Labor Law	0.54	0.54	0.56	0.57	0.74	0.82	0.85
Criminal Procedure Law	0.53	0.54	0.58	0.59	0.71	0.81	0.84
Electoral Law	0.48	0.51	0.50	0.53	0.83	0.83	0.86
Criminal Law	0.50	0.51	0.54	0.54	0.65	0.82	0.85

Table 1: Accuracy by legal domain and FIRAC stage. Cell color ranges from red (low accuracy) to green (high accuracy), with mid-range values conservatively biased toward red.

E Principal Component Anaysis

We report in Table 3 the component loadings obtained from the Principal Component Analysis (PCA); here, *unstructured* refers to prompts in which the FIRAC chain-of-thought structure is not imposed.

F Jurema-7B hallucinations

Here, we conduct a qualitative analysis of `Jurema-7B` errors by filtering questions for which `Qwen-7B-Instruct` and `Qwen-14B-Instruct` exhibit perfect monotonic trajectories, while `Jurema-7B` presents nonzero instability ($\mathrm{S2E} > 0$). Using the taxonomy of legal hallucinations proposed by (Charlotin, 2026), we classify errors into fabricated cases, misrepresented norms, and false quotes, and report all instances in which explicitly providing the correct statutory reference (`L`) unexpectedly misguides the model; results are presented as one table per legal domain. Misrepresented norms are by far the biggest offender.

Model	Accuracy	E2S	S2E
gpt-5.2-high	0.95	0.129 [0.058, 0.200]	0.112 [0.052, 0.172]
gpt-5.2-medium	0.93	0.156 [0.078, 0.233]	0.101 [0.050, 0.153]
gpt-5.2-low	0.92	0.213 [0.125, 0.301]	0.129 [0.069, 0.189]
gpt-5.2-none	0.87	0.359 [0.270, 0.449]	0.209 [0.153, 0.265]
sabia-3.1	0.86	0.441 [0.396, 0.487]	0.285 [0.254, 0.316]
sabia-3	0.74	0.863 [0.804, 0.923]	0.283 [0.256, 0.309]
gpt-5-mini-low	0.73	0.876 [0.671, 1.081]	0.436 [0.312, 0.560]
sabiazinho-3	0.73	1.029 [0.959, 1.099]	0.284 [0.258, 0.309]
Jurema-7B	0.59	1.151 [1.086, 1.215]	1.048 [1.000, 1.096]
gemma-3-27b-it	0.60	1.199 [1.137, 1.260]	0.417 [0.388, 0.446]
gpt-5-nano	0.61	1.288 [1.220, 1.356]	0.714 [0.667, 0.760]
gemma-3-12b-it	0.56	1.367 [1.302, 1.432]	0.437 [0.409, 0.465]
Qwen2.5-14B-Instruct	0.57	1.445 [1.366, 1.524]	0.452 [0.419, 0.485]
gpt-4o-mini	0.61	1.517 [1.355, 1.679]	0.434 [0.362, 0.506]
gemma-3n-e4b-it	0.51	1.541 [1.394, 1.687]	0.828 [0.748, 0.908]
gemma-3-4b-it	0.44	1.902 [1.826, 1.978]	0.877 [0.833, 0.921]
Qwen2.5-7B-Instruct	0.49	1.930 [1.844, 2.016]	0.782 [0.740, 0.825]

Table 2: Accuracy and transition-based diagnostics (E2S and S2E), reported as mean with 95% confidence intervals in brackets.

Table 3: PCA loadings for the first two principal components. Green indicates positive contributions, red negative contributions, and near-zero values are shown in white.

FIRAC level	PC1	PC2
unstructured	0.38	-0.22
_____	0.39	-0.23
F____	0.39	-0.23
FI___	0.39	-0.19
FIL__	0.39	-0.18
FIR__	0.33	0.27
FIRA_	0.28	0.56
FIRAC	0.25	0.63

Case ID	Legal Domain	Hallucination	Error Description	Erroneous Excerpt (LLM)
oab-143-250	Administrative Law	Misrepresented	The model misapplied Law 8,987/1995 by focusing on the post-intervention return of management and wrongly stating that no accounting is required, contrary to Art. 34. It also omitted the decisive rule of Art. 33, which requires opening an administrative proceeding within 30 days with full adversarial proceedings and full defense, leading to an incorrect choice.	"The administration of the service, after the intervention ends and if the concession is not terminated, will be returned to the concessionaire, regardless of the intervener's rendering of accounts, because the intervener is not responsible for the acts performed during the measure."
oab-127-084	Administrative Law	Misrepresented	The model misrepresented Decree-Law 25/1937 by treating listed private property as practically inalienable (only transferable to a public entity), when the rule only makes publicly owned listed assets inalienable and merely restricts private alienation. This distorted rule led it to mark option B as the exception, even though the real exception is the alleged duty to provide the property for use as public offices (which has no legal basis).	"Owners cannot alienate the assets, except for transfer to a public entity. ... Therefore, option D is the exception ... The correct answer is option B."
oab-138-193	Administrative Law	Misrepresented	The model misrepresented Art. 17 of Law 8,666/93 by treating it as authorizing the sale of a public-use asset by decree and without bidding based on public interest/loss reversal. It then applied this distorted rule to validate option A, ignoring the required prior disaffection and the mandatory bidding procedure for alienating an affected public asset (public-use good).	"Art. 17 of Law No. 8,666/93 allows the alienation of public assets ... and, in some cases, without bidding, when there is public interest."
oab-128-102	Administrative Law	Misrepresented	The model distorted the constitutional rule by treating a state-owned public company that explores economic activity as subject to a statutory (civil-service) regime and by concluding that a public competitive examination is dispensable. This contradicts CF/88 art. 173 §1º II (private-law labor regime) and art. 37 II (mandatory competitive examination for public jobs).	"The public company that explores economic activity is not subject to the legal regime of private companies, but rather to the statutory regime, which dispenses with the requirement of a public competitive examination for hiring its employees."
oab-131-130	Administrative Law	Misrepresented	The model misstates Art. 1 of Law 12,462/2011 as if the RDC broadly authorizes faster procurement of common goods (e.g., stationery), ignoring that RDC is limited to the legally listed hypotheses. By applying this distorted rule, it concludes RDC is appropriate and rejects the correct solution (SRP processed via pregão) for recurring purchases.	"The Federal Administration may use the Differentiated Regime of Public Contracting (RDC) under Art. 1 of Law No. 12,462/2011, which allows the contracting of common goods and services in a more agile and efficient manner... Therefore, the use of the RDC is adequate for contracting common goods and services, such as stationery products."

Table 4: Hallucination cases identified in the evaluation for the legal domain of Administrative Law for model `Jurema-7B`.

Case ID	Legal Domain	Hallucination	Error Description	Erroneous Excerpt (LLM)
oab-216-179	Business Law	Misrepresented	The model distorted Civil Code art. 1,156 by treating it as allowing a fantasy name for an individual firm and by denying that the rule permits adding an indication of the business activity. This misreading led it to pick option B and to reject the correct option C, which accurately reflects the statute.	"The article 1,156 of the Civil Code establishes that the firm must be constituted by the entrepreneur's abbreviated name or by a fantasy name..."
oab-199-012	Business Law	Misrepresented	The model misrepresented Law 9.514/1997 by treating fiduciary sale as restricted to legal entities within the SFI, contrary to Art. 22 §1, and it failed to select the Art. 23 rule that fiduciary ownership is constituted by registration. This distortion led it to choose option D instead of the correct option B.	"It may only be contracted by a legal entity that is part of the Real Estate Financing System SFI."

Table 5: Hallucination cases identified in the evaluation for the legal domain of Business Law for model `Jurema-7B`.

Case ID	Legal Domain	Hallucination	Error Description	Erroneous Excerpt (LLM)
oab-181-291	Civil Law	Misrepresented	The model mischaracterized the contractual daily penalty clause (cláusula penal moratória) as judicial astreintes and concluded it cannot be equitably reduced, contradicting Civil Code art. 413. It also ignored art. 408 by failing to recognize that the penalty is payable upon simple delay, without proof of damages, leading to the wrong option.	"The obligation to pay the fine in case of delay has the legal nature of astreintes and cannot be equitably reduced by the judge if it proves manifestly excessive."
oab-177-252	Civil Law	Misrepresented	The model mischaracterized the employers liability under Civil Code art. 932(III) as subsidiary, even though the applicable rule imposes strict (objective) direct liability with a right of recourse under art. 934. This led it to conclude that the workshop would only pay if the mechanic lacked assets, contradicting the reference FIRAC.	"Carro Bom workshop will be subsidiarily liable for the damages caused by Henrique, if he does not have sufficient assets to satisfy enforcement."

Table 6: Hallucination cases identified in the evaluation for the legal domain of Civil Law for model `Jurema-7B`.

Case ID	Legal Domain	Hallucination	Error Description	Erroneous Excerpt (LLM)
oab-52-055	Constitutional Law	Misrepresented	The model replaced the constitutional rule of concurrent legislative competence (CF art. 24, §4) with a generic more protective environmental rule prevails criterion. This misapplication led it to conclude that the stricter norm should apply even when it conflicts with a supervening federal general law, instead of recognizing the suspension of the conflicting state provisions.	"Therefore, the law that establishes more restrictive standards will be applied, regardless of whether that rule is federal or state."

Table 7: Hallucination cases identified in the evaluation for the legal domain of Constitutional Law for model `Jurema-7B`.

Case ID	Legal Domain	Hallucination	Error Description	Erroneous Excerpt (LLM)
oab-197-069	Consumer Law	False Quotes	The model wrongly attributes to CDC Art. 25(1) a rule about professionals personal liability requiring proof of fault, and then uses this distorted rule to deny joint-and-several liability in the supply chain. As a result, it incorrectly concludes that the points club is a mere intermediary with no passive standing, contrary to the CDC regime of objective and solidary liability for service defects.	"Pratice Ltda. is not responsible for providing the service, as it acts only as an intermediary between the hotels and the purchasers of the points-club title... Art. 25, §1, of the CDC specifies that the personal liability of liberal professionals will be determined upon verification of fault... Pratice Ltda. is not liable for the damages suffered by the consumer, since there is no direct consumer relationship between the consumer and the company, but rather between the consumer and the affiliated hotel."
oab-196-064	Consumer Law	Misrepresented	The model distorted CDC Art. 37, §2 by treating the easy identification of the marketing message as a factor that prevents abusiveness. This led it to conclude that the ad is only abusive when children cannot recognize it as advertising, contradicting the rule that abusiveness stems from exploiting childrens lack of judgment/immaturity regardless of recognizability.	"Advertising cannot be considered abusive or misleading if the target audience, in an easy and immediate way, identifies the marketing message as such."

Table 8: Hallucination cases identified in the evaluation for the legal domain of Consumer Law for model `Jurema-7B`.

Case ID	Legal Domain	Hallucination	Error Description	Erroneous Excerpt (LLM)
oab-247-018	Criminal Law	Misrepresented	The model distorted the Superior Courts approach to the crime of customs evasion (descaminho) by treating harm to the Treasury as automatically relevant, thereby excluding the insignificance principle. It ignored that, given the referenced STJ/STF thresholds and the defendants favorable circumstances, the correct conclusion is to allow arguing insignificance (bagatelle).	"Therefore, the insignificance principle does not apply to the crime of customs evasion (descaminho), even if the amount is low, because the damage to the Treasury is relevant."

Table 9: Hallucination cases identified in the evaluation for the legal domain of Criminal Law for model `Jurema-7B`.

Case ID	Legal Domain	Hallucination	Error Description	Erroneous Excerpt (LLM)
oab-151-077	Environmental Law	Misrepresented	The model correctly described the legal rule that the permit issued at the preliminary planning stage (approving location/conception and setting conditions) is a Preliminary License, but it inconsistently selected the alternative for an Operation License. This is an internal reasoning-to-answer mismatch that leads to a wrong conclusion despite the correct rule being available.	"A": "Operation License" ... "answer": "C"

Table 10: Hallucination cases identified in the evaluation for the legal domain of Environmental Law for model `Jurema-7B`.

Case ID	Legal Domain	Hallucination	Error Description	Erroneous Excerpt (LLM)
oab-86-099	Human Rights	Misrepresented	The model distorted the legal rule (Art. 68 ADCT and Decree 4,887/2003) by treating quilombola remnants as the Afro-descendant population in general affected by socioeconomic deprivation and discrimination. It therefore applied an overbroad concept and concluded option B, ignoring the required elements of self-identification, specific territorial ties, and a distinct historical trajectory of resistance.	"Quilombola remnants are defined as the Brazilian Afro-descendant population that undergoes processes of rights deprivation and seeks reparations for prejudice and discrimination, resulting in restrictions on access to land and housing."
oab-81-054	Human Rights	False Quotes	The model denied the criminal typification and treated the case as merely a principles-balancing/moral-damages discussion, ignoring that inciting discrimination against a person with a disability is a statutory crime when broadcast by mass media. It also attributed incorrect wording to CF/88 Art. 5(IV)-(V) and mischaracterized the content of LBI Art. 88, leading to an incorrect conclusion that there was no criminal offense.	"The case is a typical clash of principles in which, on one side, is the principle of human dignity and, on the other, the principle of freedom of expression. But there is no characterization of a civil tort nor a criminal offense."

Table 11: Hallucination cases identified in the evaluation for the legal domain of Human Rights for model `Jurema-7B`.

Case ID	Legal Domain	Hallucination	Error Description	Erroneous Excerpt (LLM)
oab-95-075	International Law	Misrepresented	The model correctly cited CPC/2015 art. 21(III) but misapplied it by treating the victims Brazilian nationality as a determining jurisdictional criterion. This distortion led it to select option A instead of grounding jurisdiction exclusively on the fact that the harmful event occurred in Brazil (option C).	"Brazilian courts are competent to hear the damages action because the accident occurred in national territory and the victim is Brazilian."

Table 12: Hallucination cases identified in the evaluation for the legal domain of International Law for model `Jurema-7B`.

Case ID	Legal Domain	Hallucination	Error Description	Erroneous Excerpt (LLM)
oab-322-251	Labor Law	Misrepresented	The model misapplied CLT Art. 10-A by treating fraud as grounds for the former partners subsidiary liability, when the sole paragraph makes fraud trigger joint and several liability with the company and current partners. This distortion led it to select option A instead of recognizing the solidary (joint and several) responsibility required by the rule.	"The former partner, due to the fraud committed, will have subsidiary liability in relation to the current partners."

Table 13: Hallucination cases identified in the evaluation for the legal domain of Labor Law for model `Jurema-7B`.

Case ID	Legal Domain	Hallucination	Error Description	Erroneous Excerpt (LLM)
oab-331-089	Labor Procedure	Misrepresented	The model distorted TST Precedent 418 and CLT art. 764(3) by treating judicial approval of the settlement as mandatory, when it is discretionary. It also misstates the challenge mechanism by suggesting an annulment action, contrary to TST Precedent 259 (rescissory action only).	"The judge has the obligation to approve the settlement, if this is the parties' legitimate will, without defects or doubts."

Table 14: Hallucination cases identified in the evaluation for the legal domain of Labor Procedure for model `Jurema-7B`.

Case ID	Legal Domain	Hallucination	Error Description	Erroneous Excerpt (LLM)
oab-46-068	Philosophy	Fabricated	The model invents non-Aristotelian degenerate regimes (kleptocracy, parliamentarianism, agoraphobia) even though the rule provided identifies tyranny, oligarchy, and democracy as the correct degenerations. By applying these fabricated categories, it selects option D and reaches a conclusion inconsistent with Aristotles typology.	"Cleptocracy, which tolerates the embezzlement by those who govern; parliamentarianism, which weakens the rulers power; and agoraphobia, which weakens citizens power."

Table 15: Hallucination cases identified in the evaluation for the legal domain of Philosophy for model `Jurema-7B`.

Case ID	Legal Domain	Hallucination	Error Description	Erroneous Excerpt (LLM)
oab-5-048	Professional Ethics	Misrepresented	The model inverted the meaning of EAOAB art. 15, §4 by treating it as a ban on lawyers participating in firms located in different sectional jurisdictions, when the rule only restricts joining more than one firm within the same sectional area. This misapplication led it to reject option C and incorrectly select option A.	"The participation of lawyers in firms headquartered in territorial areas of different sectionals is not permitted, making this alternative incorrect."
oab-2-012	Professional Ethics	Misrepresented	The model distorted the cited rules by turning the lawyers duty of independence (EAOAB art. 7(I) and CED art. 2, sole paragraph, II) into an obligation to follow the clients instructions and accept co-counsel. It also failed to apply CED/2015 art. 24, which expressly states the lawyer is not required to accept the clients imposition to work jointly with other lawyers, leading to the wrong option.	"The lawyer, upon accepting the mandate, must follow the clients instructions, especially when it involves an indication of a lawyer for joint representation."
oab-11-109	Professional Ethics	Misrepresented	The model distorted Art. 4, sole paragraph, of the OAB Code of Ethics (CED/2015), which authorizes the lawyer to refuse representation when the right at issue also applies to her. It wrongly applied this by treating delegation to another in-house lawyer as a solution, even though the whole team shares the same interest and the correct duty is to inform the company and recuse herself.	"Indicate a lawyer from her team to carry out the defense."

Table 16: Hallucination cases identified in the evaluation for the legal domain of Professional Ethics for model `Jurema-7B`.

G FIRAC Examples

Largest PC1 (overall difficulty) – Legal Philosophy

One of the most important issues in the Philosophy of Law concerns the procedure that defines a legal norm as valid. For the legal philosopher Herbert Hart, in *The Concept of Law*, the foundation of the validity of law is based on the existence of a rule of recognition, without which the existence of legal systems would not be possible. According to Hart, indicate the option that defines the rule of recognition.

A) A rule that requires human beings to perform or refrain from performing certain acts, whether they wish to or not.
B) A rule that imposes duties on all those who are recognized as citizens under the protection of the State.
C) **A rule that establishes the criteria according to which a society considers the existence of its own legal norms to be valid.**
D) A rule that recognizes excluded groups and social minorities as holders of fundamental rights.

FIRAC annotation

Facts: ["The definition of the procedure that validates a legal norm is a matter of great importance in the Philosophy of Law.", "The legal philosopher Herbert Hart, in his work 'The Concept of Law', proposes that the foundation of the validity of law is based on the existence of a rule of recognition.", "According to Hart, the absence of a rule of recognition would prevent the existence of legal systems."]

Issue: What is the definition of a 'rule of recognition' according to Herbert Hart's theory?

Rule: doctrine – Theory of the Rule of Recognition (Herbert Hart – The Concept of Law): The rule of recognition establishes the criteria and aspects by which a society considers the existence of its own legal norms to be valid, and may include obedience to moral principles or substantive values of the community.

Conclusion: The rule of recognition, from Herbert Hart's perspective, is the principle that establishes the criteria by which a society determines the validity and existence of its own legal norms.

Smallest PC1 (overall difficulty) – Labor Procedure Law

Caio, a metalworker, filed a labor lawsuit against the company Ômega seeking reinstatement to his job, because, according to his allegations, he had been dismissed during the period of union stability. He also requested the granting of a preliminary injunction aiming at such reinstatement until the end of the proceedings, based on Art. 659, X, of the CLT. (...) In this regard, indicate the correct alternative.

A) The legal nature of the decision denying the injunction is that of an interlocutory decision, and no immediate appeal is available; therefore, a writ of mandamus is admissible.
B) The legal nature of the decision denying the injunction is that of a terminating decision, and an ordinary appeal is available; therefore, a writ of mandamus is inadmissible due to the existence of an appropriate appeal.
C) **The legal nature of the decision denying the injunction is that of an interlocutory decision, no immediate appeal is available, and the injunction should have been granted.**
D) The legal nature of the decision denying the injunction is that of a final decision, and a writ of mandamus is admissible, since there is no appropriate appeal in the case.

FIRAC annotation

Facts: ["Caio, a metalworker, filed a labor lawsuit against the company Ômega.", "He sought reinstatement to his job, alleging that he had been dismissed during a period of union stability." (...) "The judge decided to deny the preliminary injunction.", "The judge allowed the proceedings to continue after denying the injunction."]

Issue: What is the legal nature of the decision that denies the preliminary injunction in labor proceedings, what is the appropriate procedural avenue to challenge it immediately, and whether the injunction for reinstatement should have been granted in the context of union stability?

Rule: CLT – Art. 893, §1 of the CLT: Interlocutory decisions rendered in labor proceedings do not allow immediate appeal, that is, they cannot be challenged by appeal instantaneously. CLT – Art. 659, X of the CLT: Authorizes the granting (..) without just cause.

Conclusion: The decision that denied the preliminary injunction has an interlocutory nature and is therefore not immediately appealable in labor proceedings. However, due to the employee's union stability, the preliminary injunction for reinstatement should have been granted.

Largest PC2 (demand for normative guidance) – Civil Procedure Law

Maria and Pedro, defendants in an action proceeding under the summary procedure, are notified, through their respective attorneys, of the judgment granting the claim. On the 23rd day following notification, Maria files an appeal. Considering the criteria regarding timeliness and effects, it is correct to state that the appeal will be

A) declared untimely by the court of first instance, which will refrain from notifying the appellee to present counterarguments.
B) **admitted as timely and, as a rule, received with both devolutive and suspensive effects, given the nature of the appeal, subject to legal exceptions.**
C) received only with devolutive effect, since granting both effects is inadmissible for the appeal in question, even if timely.
D) dismissed as untimely, with the decision being rendered by the appellate court.

FIRAC annotation

Facts: ["Maria and Pedro are defendants in a judicial action.", "The action proceeds under the summary procedure.", "They were notified of the judgment that granted the claim.", "Each defendant is represented by his or her own attorney.", "On the 23rd day after notification, Maria filed an appeal."]

Issue: Is the appeal filed by Maria timely, and what effects should it have?

Rule: Code of Civil Procedure – Art. 229 of the Code of Civil Procedure: Grants a doubled deadline for joint litigants represented by attorneys from different law firms. Code of Civil Procedure – Art. 1.003, §5 of the Code of Civil Procedure: Establishes a 15-day deadline for filing an appeal. Code of Civil Procedure – Art. 1.012 of the Code of Civil Procedure: Provides that appeals are, as a rule, granted both devolutive and suspensive effects, subject to the exceptions set forth in §1.

Conclusion: The appeal filed by Maria should be admitted as timely and, as a rule, received with both devolutive and suspensive effects, subject to legal exceptions.

Smallest PC2 (demand for normative guidance) – Professional Ethics

Marcelo, a lawyer, is accused of using a false medical certificate to secure his client's release from prison. The matter gained significant public attention, to the point that a local newspaper published an article stating that Marcelo should be preventively suspended by the Brazilian Bar Association (OAB) until the disciplinary investigation of his conduct is concluded. On this topic, indicate the correct statement.

A) It is incumbent upon the Ethics and Discipline Tribunal of the Sectional Council before which the infraction occurred to suspend him preventively.
B) **Preventive suspension presupposes a demonstration that the fact has generated repercussions detrimental to the dignity of the legal profession.**
C) If preventive suspension is applied, the disciplinary proceeding must be concluded within a maximum period of sixty days.
D) Before preventive suspension is applied, the accused must be heard in a special session, unless it is not possible to notify him to appear.

FIRAC annotation

Facts: ["A lawyer is accused of using a false medical certificate to secure the release of his client from prison.", "The fact generated significant public attention.", "A local newspaper published an article stating that the lawyer should be preventively suspended by the Brazilian Bar Association while the disciplinary investigation of his conduct is ongoing."]

Issue: What are the conditions and the competent body for the preventive suspension of a lawyer by the Brazilian Bar Association, and what procedures and deadlines apply to this measure?

Rule: OAB Statute – Art. 70, §3, EAOAB: The Ethics and Discipline Tribunal of the Council where the accused has his or her principal registration is responsible for preventive suspension. Suspension may occur in cases of repercussions detrimental to the dignity of the legal profession. The Tribunal may suspend the accused after hearing him or her in a special session for which he or she must be notified to appear, unless he or she fails to respond to the notification. If preventive suspension is applied, the disciplinary proceeding must be concluded within a maximum period of ninety days.

Conclusion: The preventive suspension of a lawyer is a prerogative of the Ethics and Discipline Tribunal of the Council of the principal registration, applicable in cases of repercussions detrimental to the dignity of the legal profession. This measure requires prior hearing of the lawyer in a special session, unless he or she fails to appear after notification. Furthermore, the disciplinary proceeding associated with preventive suspension must be concluded within 90 days.