From Recommendation to Deliberation: Explainable AI for Judicial Decision Support in Brazil

Eduardo Caruso Barbosa Pacheco[1]

[1]*Institute of Mathematical and Computer Sciences (ICMC), University of São Paulo, São Carlos, Brazil*

Abstract

The Brazilian judiciary is adopting Generative AI faster than it scrutinises the capabilities of the models it deploys: 45.8% of courts already use such tools, chiefly in text-related tasks. The dominant paradigm treats model output as a *recommendation* to be accepted or rejected, rather than as material to be *deliberated* upon, which is in tension with the constitutional duty to give reasons (Art. 93, IX, of the Federal Constitution) and with CNJ Resolution 615/2025. This proposal argues that AI-assisted judicial decision-making should be redesigned around a *deliberative* paradigm: instead of justifying a single recommended outcome, the AI organises the plural space of legally admissible grounds for *opposing* outcomes, leaving the choice of outcome and its motivation to the judge. The programme articulates three interdependent pillars: *explanation*, with an ontology, a knowledge graph, and contrastive retrieval; *evaluation*, with a benchmark of the written tasks of the magistracy calibrated by LLM-as-a-judge against a human gold standard; and *training*, investigating what constitutes legal competence in Portuguese-language language models. The empirical design compares five explanation protocols pairwise, measuring perception through Bradley–Terry, the variability of the theses actually adopted, and the deliberative quality of the written justification. Published preliminary results reach a Cohen's κ of 0.79 and show that the best LLM judges are not the best respondents.

Keywords

Explainable AI, Contrastive explanation, AI and Law, LLM-as-a-judge, Legal benchmark, Judicial autonomy

1. Context and motivation

The Brazilian judiciary is one of the largest and most digitised in the world, and it has rapidly adopted Artificial Intelligence to cope with its case load. The *Survey on Artificial Intelligence in the Judiciary 2024*, by the National Council of Justice (CNJ), identified 178 projects registered in 2024, of which 98 were new initiatives, with participation from 92 of the country's 95 courts and councils (96.8% adhesion); 58 of them (63%) developed AI projects that year. For the first time the survey dedicated a specific section to mapping Generative AI, revealing that 45.8% of courts and councils already use these tools, mainly in text-related tasks (generation, refinement, and summarisation), and that, among those that had not yet adopted them, 81.3% expressed the intention to do so.[1] Despite this scale, adoption has to a large extent outpaced the scientific scrutiny of the models' capabilities.

The dominant interaction paradigm treats AI output as a *recommendation* to be accepted or rejected, rather than as material to be *deliberated* upon. This last point runs against the constitutional duty to give reasons (which requires that judicial decisions be grounded in legally admissible arguments, Art. 93, IX, of the Federal Constitution) and against CNJ Resolution 615/2025, which frames AI in courts as an instrument subject to human supervision, contestability, and non-delegation of judicial authority. The current adoption pattern conflates *predictive performance* with *judicial competence*, and *output* with *explanation*. The dominant pipeline, in which the LLM produces a draft and a human signs it, risks reducing the judge to a notary of the model. It also rests on a problematic implicit assumption: that the model is omniscient and objective, when in practice many design choices (data, evaluator,

Late-breaking work, Demos and Doctoral Consortium, colocated with the 4th World Conference on eXplainable Artificial Intelligence: July 01–03, 2026, Fortaleza, Brazil

EMAIL: eduardo.caruso.pacheco@usp.br (E. C. B. Pacheco)

ORCID: 0009-0007-9887-9189 (E. C. B. Pacheco)

[1]Conselho Nacional de Justiça, *Pesquisa Inteligência Artificial no Poder Judiciário 2024: resumo executivo*, CNJ, Brasília, 2025. https://www.cnj.jus.br/wp-content/uploads/2025/09/relatorio-ia-2024-resumo-executivo-set2025.pdf (accessed 1 August 2026).

training, prompting) remain hidden. The result is a tacit transfer of normative authority to opaque artefacts. Addressing this gap requires a paradigm shift in how XAI is conceived for judicial settings: rather than ranking, explaining, or justifying a single recommended outcome, the AI should expose the plural, legally admissible space of justificatory grounds within which the judge exercises discretion.

2. Related work

XAI for legal decisions. Case-based approaches [8] explain decisions by retrieving similar precedents, a strategy that aligns with common-law adjudication but maps poorly onto civil-law systems, in which statutes, principles, and doctrine dominate justification. Argumentation-based approaches, grounded in abstract argumentation [9] and refined by procedural models such as Carneades [10], formalise defeasible reasoning and burdens of proof; they tend, however, to collapse deliberation into a single resolved justificatory chain. Both strands underrepresent the explicit pluralism of admissible reasoning that characterises Brazilian judicial practice under the constitutional duty to give reasons (Art. 93, IX, of the Federal Constitution). Contemporary XAI emphasises that explanations are social, pragmatic, and contrastive artefacts [2, 3], sensitive to the audience and to a contrast class. The civil-law setting sharpens this requirement: judges require a deliberative view of admissible grounds; parties require an outcome-centred motivated justification; institutional auditors require traceability over both. LGPD Article 20 reinforces transparency and human oversight in automated decisions; CNJ Resolution 615/2025 frames AI in courts as auxiliary, supervised, auditable, and bound by due process, a regulatory layer rarely engaged by the technical literature. Existing approaches are inadequate to the Brazilian context because they do not respect the constitutional mandate to give reasons or the norms of CNJ Resolution 615/2025: they are mechanical winner-selection procedures (over cases or theses) and thereby lead to a usurpation of judicial competence. The measurement of deliberative quality has a consolidated instrument in deliberative-democracy theory, the Discourse Quality Index (DQI), covering justification, respect, and counter-arguments; recent work integrates it with abstract argumentation via ranking-based semantics that assign graded strength to arguments [6]. We adopt it in a distinct function: not to elect winning arguments, but to gauge the quality of the human justification produced under each protocol, as a metric rather than a decider.

LLM-as-a-judge in legal evaluation. LLM-as-a-judge methods have become the de facto standard for scaling the evaluation of free-form model outputs [11], including in legal settings, where automated writing assessment has been validated against expert reviewers [12]. The Rabula benchmark, developed by the candidate, adapts this paradigm to the second phase of the OAB Bar Exam, covering four task families and validating LLM judges against a gold standard of three human experts [1]. Even so, most published Portuguese legal benchmarks rely on multiple-choice tests, underrepresenting the argumentative competences relevant to judicial work.

Adaptation of LLMs to Brazilian Law. Existing Brazilian legal LLMs (Sabiá, Juru, Tucano) report performance mainly on multiple-choice tasks [1], with limited evidence on generative competence; whether continued pre-training, supervised fine-tuning, or distillation produce real improvements in legal reasoning (or merely shift evaluator agreement) remains underexplored.

3. Objectives

The proposal pursues four objectives: (O1) to study whether the shift to a *Decision-Space Organisation* paradigm improves utility, reduces bias, and raises the deliberative quality of the justification produced by judges; (O2) to deliver a community benchmark for Portuguese legal LLMs covering the written tasks of the magistracy, together with a benchmark for good representations of legal cases; (O3) to deliver a trained Portuguese legal LLM, together with a characterisation of what defines respondent competence; and (O4) to deliver a contrastive deliberative XAI framework, with an ontology, a knowledge graph, and contrastive retrieval.

4. Research questions and hypotheses

The thesis holds that AI-mediated judicial decision support in Brazil should be redesigned around a *deliberative* rather than a *recommendation* paradigm: moving from systems that output a single decision plus an explanation to systems that organise and expose a structured space of legally admissible grounds for opposing outcomes, leaving the selection of the outcome and its motivation to the human judge.

The thesis has an empirical stage and an AI-engineering stage. In the empirical stage we test explanation protocols for representing cases and their decisions, assessing the judges' perception of the limits to their autonomy (PE1) and of each protocol's utility (PE2), the deliberative quality of the written justification via the Discourse Quality Index adapted to monological judicial reasoning [6] (QE1), and their effective choices (AE1). Materials are curated by human experts, so as to test the structure in the abstract and separate adequacy-in-thesis from implementation. For each experiment we state the directional *alternative*, the null being its negation.

4.1. Empirical stage

Exp. 1-A. Is Decision-Space Organisation (DSO) perceived as less inducing than the other protocols?
In pairwise comparisons of the pressure felt to adopt a specific output, the latent inducement score of DSO (Bradley–Terry, maximum likelihood) is lower than that of Argument-Based Reasoning (ABR): $\beta_{\text{ind}}(\text{DSO}) < \beta_{\text{ind}}(\text{ABR})$, with twin hypotheses against each of the remaining protocols.

Exp. 1-B. Is DSO perceived as more useful than the other protocols?
Its latent utility score is higher than that of ABR: $\beta_{\text{util}}(\text{DSO}) > \beta_{\text{util}}(\text{ABR})$.

Exp. 1-C. Do protocols that deliver a winning thesis reduce the variability of the theses adopted?
The convergence rate between judges under the winner-delivering protocols (free LLM use, case-based reasoning, argument-based reasoning) is higher than under the no-AI protocol.

Exp. 1-D. Does DSO reduce convergence relative to the protocol currently in use?
The convergence rate under DSO is lower than under free LLM use, $\lambda(\text{DSO}) < \lambda(\text{P1})$, indicating greater variability of the theses adopted.

Exp. 1-E. Does DSO raise the deliberative quality of the justification produced by the judge?
Coding the written justification by the monological DQI indicators [6] (level and content of justification, counter-arguments), the mean score under DSO is higher than under free LLM use; the comparison with the no-AI protocol tests amplification, since justification above the unassisted judge means the system amplifies human reasoning, while justification below it under free LLM use means the recommendation paradigm degrades it.

4.2. AI-engineering stage

In the AI-engineering stage we work on three fronts. On front **EIA1** we build specific benchmarks to evaluate the competence of specialised legal models both in the tasks proper to judges (B1), which coincide with the abilities assessed in the magistracy entrance exam (legal opinion, legal reflection, and sentence writing), and in the ability to represent cases (B2).

Exp. 2-A. Can we reliably and automatically measure the competence of AI models in judicial tasks?
With a ruler of three judges or assessors judging, criterion by criterion, whether respondent-model answers on magistracy entrance exams are acceptable, and LLMs in the role of evaluators, Cohen's κ between models and humans exceeds 0.75 in each of the three tasks (legal opinion, legal reflection, and sentence writing).

Exp. 2-B. Can we reliably and automatically measure the competence of AI models in representing judicial cases?
Under the same ruler over representations of real TJSP cases, Cohen's κ exceeds 0.75 in the elements that compose the case representation.

On front **EIA2** we work on building models that are competent both in the tasks proper to judges and in the representation of judicial cases.

Exp. 3-A. Can a compact model be endowed with competence in the tasks proper to the magistracy?
Applying steps (a)–(c) (adapted tokenizer, continued pre-training, sequence-level distillation) to an open base model of at most 21 billion parameters and measuring it with benchmark 1 under an evaluator with $\kappa > 0.75$, the median score across areas of law exceeds 70 in each of the three tasks.

Exp. 3-B. Can a compact model be endowed with competence in representing judicial cases?
Applying step (d) to the 3-A model and measuring it with benchmark 2 under an evaluator with $\kappa > 0.75$, the proportion of representation elements evaluated as adequate exceeds 85%.

On front **EIA3** we produce the mechanism that generates the organisation of the decision space with contrastive theses. This proceeds through the representation of past cases according to a legal ontology developed for Brazilian law and incorporated into a knowledge graph; from this, similar cases can be retrieved via GraphRAG given a pending case awaiting judgment. For this front we reproduce experiments 1-A to 1-D, but this time generated by the mechanism. We check whether the results hold: the null hypothesis is that we will not be able to reproduce the earlier results with the automatic mechanism; the alternative is that we will.

5. Approach, methods, and rationale for hypothesis testing

5.1. Pillar 1 – Explanation: navigating justificatory spaces

Motivated by van Fraassen [3] and Miller [2], and grounded in AI & Law argumentation [9, 10], this pillar develops a contrastive and deliberative XAI framework for Brazilian judicial decisions [4, 5]. A pending case is modelled as $C = (P, F^1, F^2, \Pi)$, with P the claim, F^1, F^2 the parties' arguments, and Π the procedural configuration. For each candidate outcome $O \in \{D, \neg D\}$, the system organises a set $G(C, O, \mathbf{w})$ of legally admissible justificatory grounds drawn from statutes, principles, jurisprudence, and doctrine, contextualised by the institutional support $\mathbf{w}$. An explanation is a situated selection $E = (C, O^*, g^*)$ performed by the human judge, rendered differently for judges (the full decisional space), parties (the selected decision and ground), and institutional oversight (the combination of the available possibilities and the effective choice). Operationally, the framework is realised as a retrieval-augmented generation pipeline grounded in a knowledge graph, along two paths: a retrieval conditioned on C and the candidate outcome O returns admissible grounds for O; a parallel retrieval, conditioned on C and $\neg O$ over the same case, returns admissible grounds for the opposite outcome, making the contrast *normative* rather than factual. The LLM then synthesises justificatory drafts anchored in the retrieved grounds, with provenance preserved through the graph for auditability.

A central representational device are structured legal *cards*: each statute, doctrinal entry, jurisprudential thesis, and first-instance decision is expanded into a card carrying the original text plus contextual paraphrases, syntheses, scope-of-application cues, typical practical effects, and cross-references to related sources. This enrichment opens *multiple semantic access routes* to the same legal source, so that a pending case retrieves it even when its literal wording diverges, a key requirement in civil-law adjudication, in which the same norm is invoked under many linguistic surfaces. Doctoral extensions: a domain ontology for Brazilian legal sources; the knowledge graph linking statutes, jurisprudence, doctrine, and adjudicated cases through these cards; and ranking schemes that score the candidates in

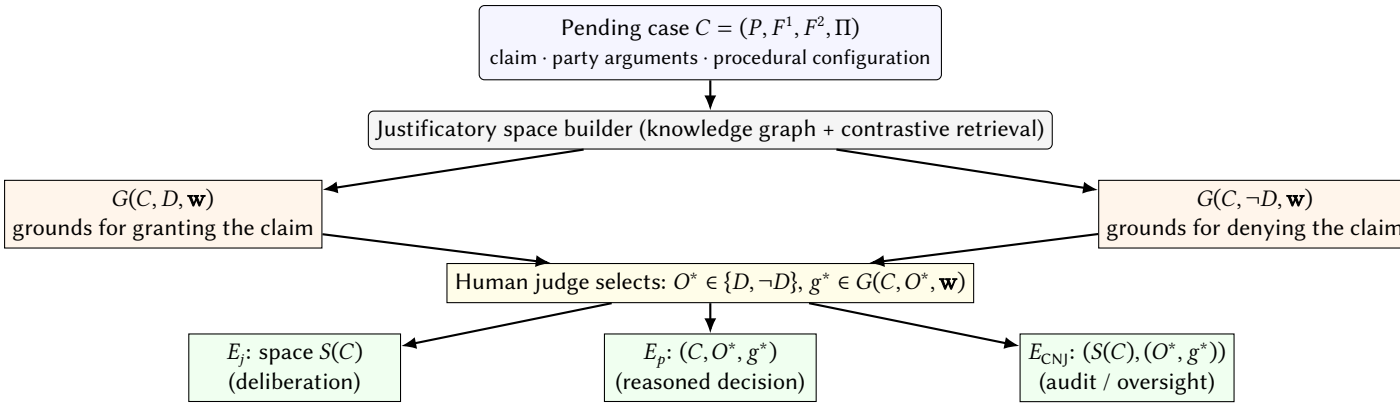

Figure 1: Deliberative XAI framework. A pending case is mapped to a plural decisional space; the human judge selects the outcome and the ground.

$G(C, O, \mathbf{w})$ by normative sufficiency rather than by mere semantic similarity to the case. The framework is testable through a controlled comparison of the explanation protocols over the same set of cases, evaluated by human experts and by LLM judges (Pillar 2). This separates conceptual gains from artefacts of LLM-as-a-judge calibration. Figure 1 summarises the resulting conceptual pipeline.

5.2. Pillar 2 – Evaluation: a benchmark for judicial LLMs

The benchmark extends the Rabula methodology [1] into a community resource covering the *written* tasks that define professional practice in different Brazilian legal careers. The master's-level Rabula covered four task families of the second phase of the OAB Bar Exam (multiple choice, document selection, document drafting, and essay-style case resolution), which is the entry exam for lawyers. The doctoral extension builds a benchmark of its own for the magistracy, covering the tasks of legal opinion, legal reflection, and sentence writing. Each written exam produces a professional artefact (a legal document of a specific type) together with its official rubric. The rubric, combined with the candidate's response and a human gold standard (majority vote of three experts), is the substrate against which LLM-as-a-judge protocols are calibrated. The protocol uses LLM judges with majority voting across multiple seeds, audited via Cohen's Kappa against the human standard.

The benchmark is the operational instrument capable of falsifying claims about model competence and judge calibration. The asymmetry between judges and respondents, found in the first expansion of the Rabula benchmark, is directly testable as a 2×2 pattern on every model: ⟨score as judge, score as respondent⟩. Preliminary results already exhibit this pattern (Section 6).

5.3. Pillar 3 – Training: what defines a competent legal respondent

The driving question is not how to optimise an adapter, but *what constitutes legal competence in a Portuguese-language LLM*, given that the best LLM judges in the domain are not the best respondents, and vice versa. Training is treated as an investigative tool, in four chained steps: (a) tokenizer adaptation, (b) continued pre-training, (c) sequence-level distillation on the tasks proper to the magistracy, and (d) verifiable-reward optimisation, taking the Pillar 1 benchmark as the reward function. The corpus, already captured, gathers 6.4 million documents and 6.6 billion tokens of legislation, higher- and state-court jurisprudence, first-instance decisions, and doctrine, enriched by synthetic case briefs that organise each statutory article into blocks derived from the normative structure. Cheng et al. [15] show that pre-training on raw corpora degrades instruction-following, a degradation corrected by transforming the corpus into reading-comprehension texts; our variation lies in deriving the blocks from the norm itself rather than from generic instruction repertoires. Tokenizer adaptation precedes training, since it is there that the model learns the new vocabulary, and it answers a question of its own: frequent expressions of legal Portuguese, such as "agravo de instrumento", are today fragmented

Table 1
Mapping of pillars to research questions, central hypotheses, and primary outputs.

Pillar	RQ	Central hypothesis	Primary output
1 – Explanation	RQ1	The organisation of the decisional space $S(C)$ is perceived as more useful and less inducing than protocols that elect a winning thesis, preserves the variability of the theses adopted, and raises the deliberative quality of the justification	Ontology, knowledge graph, and contrastive-retrieval pipeline; pairwise comparison protocol
2 – Evaluation	RQ2	The written tasks proper to the magistracy allow LLM-as-a-judge to be calibrated to inter-human κ; good judges $\neq$ good respondents	Open benchmark for judicial tasks and case representation; judge ranking
3 – Training	RQ3	Legal competence is data-centric; respondent and evaluator traits are structurally distinct	Open Portuguese legal LLM, with a reproducible recipe and a characterisation of competence

into many tokens; we measure fertility before and after, sharpening how much of legal competence is vocabulary and how much is reasoning. With a cap of 21 billion parameters in fp8, the model fits infrastructure accessible to a court. The methodological output is not a trained model, but a characterisation of competent legal response, with a reproducible recipe and a published checkpoint.

The hypothesis that data-centric interventions dominate architecture-centric adaptation is testable through controlled comparisons under Pillar 2, taking the base model before any adaptation as the baseline, measured at each scale considered. The structural distinction between competent-respondent and competent-evaluator traits is testable by training models optimised for one role and measuring transfer to the other, which separates the conceptual question of *what* competence is from the operational question of *how* to obtain it.

5.4. Integration across pillars

Pillar 2 makes the claims of Pillars 1 and 3 falsifiable. Pillar 3 produces the technical substrate on which Pillar 1 (contrastive explanation) is exercised. Pillar 1 reshapes the kind of output that Pillar 3 should optimise for, displacing the target from a single recommendation to a structured justificatory space. The central thesis holds only with the three pillars articulated: a benchmark without a competent model risks measuring noise; a competent model without contrastive explanation reinstates the recommendation paradigm (Table 1).

6. Preliminary results and contributions to date

Pillar 2 – Rabula (published). The Rabula benchmark was published in AMELR 2025, colocated with ICAIL 2025 [1], with LLM-as-a-judge calibration against three experts reaching a Cohen's Kappa of up to 0.79 on discursive questions (close to the inter-human Fleiss' Kappa of 0.78–0.79). An extended version under preparation (targeting JBCS) expands to more than 20 judges and 24 respondents, introduces a hybrid configuration, and produces three central findings: (i) a statistically meaningful systematic bias of `gpt-4o-mini-solo`, a judge commonly reported in legal LLM evaluations, in favour of fine-tuned models, suggesting that part of the published improvements in legal LLMs may be evaluation artefacts; (ii) an ensemble of three judges exceeds the inter-human Fleiss' agreement (0.83 vs. 0.79 on discursive questions); and (iii) the best judges are not the best respondents: Kimi K2.5 and GPT-OSS 120B reach a Kappa above 0.78 against humans while figuring among the worst respondents, whereas the best respondent (Mistral L3) is only a mid-tier judge. Finding (iii) is the empirical motivation for treating "competent evaluator" and "competent respondent" as structurally distinct objects, a commitment that propagates to Pillar 3 (training).

Pillar 1 – Justificatory Spaces. A theoretical framework formalising explanation as the organisation of a plural decisional space, with audience-sensitive rendering, was developed and accepted at a workshop

on legal ontologies [4, 5].

Pillar 3 – Tucano-OAB (paper in preparation, JURIX 2026). An empirical study mapped which types of intervention move the Tucano [1] base model in the direction of legal respondent competence: twenty training sessions (12 continued pre-training, 8 supervised fine-tuning) yielded 14 evaluable models. The findings indicate that competence requires data-centric ingredients absent from default LoRA [14] recipes: targeting MLP layers [13], including input/output embeddings, and longer SFT contexts for document drafting. Under calibrated judges, however, the best architecture-centric variant exceeds the base by only +1.89 points, while the gap to the strongest permissively licensed respondent (Mistral L3, 71.35 vs. 26.38) remains large, suggesting that closing the gap requires curated distillation rather than further adapter tuning.

7. Expected next steps and final contribution

The doctorate is organised in three blocks. **Block A – methodological consolidation.** The two measurement instruments are built: the benchmark of the written tasks of the magistracy (legal opinion, legal reflection, and sentence writing) and the benchmark of case representation, both calibrated against a gold standard of three judge annotators, with the criteria of the official rubric atomised into binary items and LLM judges ranked by Cohen's κ. The ontology of Brazilian legal sources, already sketched in the candidate's earlier publications, is refined and populated as a knowledge graph linking statutes, jurisprudence, doctrine, and adjudicated-case cards, restricted to one branch of law as a proof of concept and published under a permissive licence. A first round of distillation with compatibly licensed teachers, under human-in-the-loop curation, generates the substrate consumed in Block B. Submission to the research ethics committee.

Block B – empirical core. A controlled study compares five explanation protocols (absence of AI, free LLM use, case-based reasoning, argument-based reasoning, and decision-space organisation) over a common set of cases, contrasted pairwise. Three families of measures test RQ1: perception of utility and of inducement, estimated by Bradley–Terry [7] with the reference fixed on the no-AI protocol; effective behaviour, through the convergence rate between judges; and the deliberative quality of the written justification, through the monological indicators of the Discourse Quality Index [6]. In this block the materials are curated by human experts, separating the adequacy of the paradigm in thesis from the quality of its implementation. In parallel, the training pipeline is executed in its four steps, with ablations isolating the effect of the adapted tokenizer and of the corpus enrichment, testing RQ2 and RQ3 under a shared evaluation budget and taking the base model before any adaptation as the baseline.

Block C – integration and dissemination. The Block B experiment is repeated with materials generated by the automatic mechanism, and the difference between the two moments measures how much of the conceptual advantage survives automation. An open platform integrates benchmark, adapted model, contrastive retrieval, and audience-sensitive rendering. As an additional goal, a pilot at an institutional partner under LGPD assesses whether contrastive deliberative XAI changes how AI output is incorporated into legal reasoning; broad deployment is out of scope. Target submissions: JURIX, ICAIL, and an extended version in *Artificial Intelligence and Law*, plus a policy note on LGPD Article 20 and CNJ Resolution 615/2025.

The expected contribution spans five axes: *conceptual* (deliberative XAI from normative contrast and stratified audiences), *methodological* (a multitask, multi-area legal benchmark calibrated by human agreement), *technical* (an open training–retrieval–rendering pipeline), *empirical* (joint measurement of perception, behaviour, and argument quality), and *governance* (requirements mapped onto LGPD and CNJ Resolution 615/2025). The originality lies in the integration: redefining what it means for AI to *support* a judicial decision under the constitutional duty to give reasons.

Declaration on Generative AI

The author employed Generative AI tools (Anthropic Claude) for language editing, Portuguese–English translation, and LaTeX formatting. All scientific content and its interpretation are the author's own, and the author reviewed and edited the entire text and takes full responsibility for it.

References

[1] E. C. B. Pacheco, F. M. Suriani, R. Ribeiro, Rabula: A benchmark for evaluating LLMs in Brazilian legal tasks, in: Proceedings of the First Argument Mining and Empirical Legal Research Workshop (AMELR 2025), CEUR Workshop Proceedings 4089, 2025, pp. 55–62. URL: https://ceur-ws.org/Vol-4089/paper6.pdf

[2] T. Miller, Explanation in artificial intelligence: Insights from the social sciences, Artificial Intelligence 267 (2019) 1–38.

[3] B. C. van Fraassen, The pragmatic theory of explanation, Studies in History and Philosophy of Science 25 (1994) 683–704.

[4] E. C. B. Pacheco, C. M. Russo, A justificatory-space ontology and knowledge graph for contrastive explanation of Brazilian judicial decisions, in: 19th Seminar on Ontology Research in Brazil (ONTOBRAS) / WTDO, CEUR Workshop Proceedings, 2026, in press.

[5] E. C. B. Pacheco, C. M. Russo, A neuro-symbolic and contrastive semantic framework for explainable judicial AI, in: Workshop on Ontology and Semantics of Explanations and Trustworthy AI (OSET-AI), JOWO, CEUR Workshop Proceedings, 2026, in press.

[6] S. Kumar, J. Suiter, L. Longo, Advancing deliberative discourse measurement: the intersection with computational abstract argumentation in discourse quality evaluations, Systems 13 (2025) 204. doi:10.3390/systems13030204.

[7] R. A. Bradley, M. E. Terry, Rank analysis of incomplete block designs: I. The method of paired comparisons, Biometrika 39 (1952) 324–345. doi:10.1093/biomet/39.3-4.324.

[8] T. J. M. Bench-Capon, G. Sartor, A model of legal reasoning with cases incorporating theories and values, Artificial Intelligence 150 (2003) 97–143.

[9] P. M. Dung, On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games, Artificial Intelligence 77 (1995) 321–357.

[10] T. F. Gordon, H. Prakken, D. Walton, The Carneades model of argument and burden of proof, Artificial Intelligence 171 (2007) 875–896.

[11] L. Zheng, W.-L. Chiang, Y. Sheng, et al., Judging LLM-as-a-judge with MT-Bench and Chatbot Arena, NeurIPS Datasets and Benchmarks Track, 2023.

[12] R. Pires, R. Malaquias Jr., R. Nogueira, Automatic legal writing evaluation of LLMs, in: 20th International Conference on Artificial Intelligence and Law (ICAIL 2025), ACM, 2025, pp. 420–424.

[13] M. Geva, R. Schuster, J. Berant, O. Levy, Transformer feed-forward layers are key-value memories, in: Proceedings of EMNLP 2021, 2021, pp. 5484–5495.

[14] E. J. Hu, Y. Shen, P. Wallis, et al., LoRA: Low-rank adaptation of large language models, in: ICLR, 2022.

[15] D. Cheng, S. Huang, F. Wei, Adapting large language models via reading comprehension, in: ICLR, 2024.