JACV: A Benchmark for Legal-Argumentative Coherence Verification in Brazilian Portuguese

Willgnner F. Santos[1,2], João P. C. Presa[1], Sávio S. T. de Oliveira[1], Antônio P. de Castro Júnior[2], Fernando R. Trindade[2], Marcella S. R. Martins[3], Paulo V. dos Santos[4], Ronaldo M. da Costa[1] and Nádia F. F. da Silva[5]

[1]Institute of Informatics – UFG (INF/UFG), Goiânia – GO, Brazil

[2]Directorate of Artificial Intelligence, Data Science and Statistics (DIACDE)
Court of Justice of the State of Goiás (TJGO), Goiânia – GO, Brazil

[3]Federal University of Technology – Paraná (UTFPR), Brazil

[4]Institute of Innovation in Management – UFG (IIG/UFG), Goiânia – GO, Brazil

[5]Federal University of Catalão (UFCAT), Catalão – GO, Brazil

{willgnner_santos,joaopaulop}@discente.ufg.br

{savioteles,ronaldocosta,nadia.felix}@ufg.br

Abstract. *We present JACV, a benchmark for legal-argumentative coherence in Brazilian Portuguese procedural documents, built from 120 valid anchors drawn from 1,000 authentic anonymized documents. JACV comprises two tasks: three-class classification (JACV-CLS) and adversarial relief selection (JACV-ADV). Across statistical baselines, encoders, fine-tuned models, LLMs, and Legal Chain-of-Thought prompting, the best models reach 99.58% on JACV-ADV and 76.94% on JACV-CLS; a TF-IDF baseline achieves 98.33% on JACV-ADV, exposing residual lexical separability, while Legal-CoT yields parsing error rates of 47–58%, revealing protocol-level limitations. Code and benchmark instances are available in a public repository.*

1. Introduction

In recent years, the growth of procedural volume in Brazil has renewed interest in Artificial Intelligence (AI) solutions applied to the Judiciary, covering tasks of triage, document organization and support for reading procedural documents [Conselho Nacional de Justiça 2025a]. At the same time, the adoption of language models in the legal domain has highlighted a central problem: satisfactory performance in general classification, retrieval or thematic labeling tasks does not necessarily equate to the ability to legally reason about the internal coherence of a document [Guha et al. 2023, Fei et al. 2024].

In the procedural context, this limitation is sensitive. A petition may be linguistically plausible and yet present incompatibility between the factual narrative, the legal reasoning and the relief sought. The literature on textual inference has shown that neural models exploit superficial lexical shortcuts rather than deeper inferential relations [McCoy et al. 2019]. In legal applications, this risk becomes critical, as systems

aimed at decision support or workflow automation cannot be validated by textual proximity alone when the core problem lies in logical-argumentative consistency.

This work proposes the *Juridical Argument Coherence Verification* (JACV), an evaluation protocol aimed at legal-argumentative coherence in Brazilian Portuguese (pt-BR). The contribution is organized along three fronts: **(A1)** the curation of an anonymized legal corpus structured around the fields fact, law and relief; **(A2)** the construction of a benchmark with original and controlled counterfactual examples, designed to test whether models can distinguish coherent, incoherent and contradictory relations between legal reasoning and relief; and **(A3)** the comparative evaluation of multiple inference strategies, covering statistical baselines, embedding-based approaches, supervised fine-tuning, zero-shot inference and legal prompting. The relevance is both methodological (proposing an evaluation standard aligned with legal reasoning beyond textual similarity) and applied, providing a rigorous criterion for assessing model reliability in judicial contexts, where legally incoherent responses may lead to significant analytical errors [Dahl et al. 2024].

2. Related Work

In Legal NLP for pt-BR, important work has focused on creating resources and specialized models. VICTOR [Luz de Araujo et al. 2020] stands among the first large-scale resources for Brazilian legal documents; LegalNLP [Polo et al. 2021] expanded this ecosystem with pre-trained models and tools for Brazilian legal language; CDJUR-BR [Brito et al. 2023] strengthened named entity recognition in the Brazilian Judiciary; and RoBERTaLexPT [Garcia et al. 2024] advanced legal pt-BR models with domain-specific pre-training, also introducing PortuLex for supervised legal task evaluation. JurisBERT [Viegas et al. 2023] proposed an STS-based legal text classification approach, and LegalBERT-pt [Silveira et al. 2023] extended the BERT pre-training paradigm to Brazilian legal corpora. BERTimbau [Souza et al. 2020] established a foundational pt-BR model used across various legal domain adaptations.

Internationally, LegalBench [Guha et al. 2023] provides a broad benchmark for legal reasoning in LLMs, LawBench [Fei et al. 2024] evaluates legal knowledge across multiple cognitive levels, and COLIEE [Goebel et al. 2024] remains a key reference for legal retrieval and entailment. ECtHR-PCR [T.y.s.s. et al. 2024] introduces precedent retrieval with explicit fact/argument separation. Recent work has also examined prompting strategies: Yu et al. [Yu et al. 2023] show that zero-shot, few-shot and explanation-based prompting substantially alter model behavior on COLIEE tasks; Servantez et al. [Servantez et al. 2024] propose explicit decomposition for rule-based legal reasoning; and Chu et al. [Chu et al. 2025] investigate how Chain-of-Thought affects interpretability and alignment in discursive legal tasks. Compared to these resources, JACV addresses specifically the argumentative coherence between legal reasoning and requested relief in Portuguese-language procedural documents, a dimension not targeted by any prior pt-BR benchmark.

3. Methodology

3.1. Corpus, Design, and Data Accounting

The study draws on a corpus of **1,000 authentic procedural documents in pt-BR**, sourced from the archives of the Court of Justice of the State of Goiás, structured around the fields

Table 1. Data accounting throughout the benchmark construction.

Stage	N
Original corpus	1,000
Eligible after quality filters	725
Sampled anchors	200
Valid anchors for benchmark	120
Final JACV-CLS instances	360
Final JACV-ADV instances	240

Table 2. Corpus statistics by textual field (tokens).

Field	Mean	Median	P95
Full text	13,214	5,244	33,447
Fact	2,662	956	10,875
Law	5,360	1,925	11,099
Relief	7,134	703	21,605

full text, fact, law, and relief. Documents were classified by procedural type (appellate vs. non-appellate) in a balanced 500/500 distribution. The experimental design comprised four stages: structural and lexical characterization, multi-layer anonymization, JACV benchmark construction via counterfactual generation and validation, and comparative evaluation across method families.

Table 1 summarizes the data funnel. Corpus reduction occurred only at the eligibility phase: of the 1,000 documents, 725 remained after minimum quality filters (at least 80 words in fact, 150 in law, and 50 in relief; at most 5,000 words in law). From these, 200 anchors were sampled with a balanced 50/50 appellate/non-appellate distribution, a deliberate design choice ensuring equal statistical power across document types and preventing coherence judgment from being confounded with document-type recognition. In the end, 120 anchors produced complete pairs of accepted perturbations, yielding 360 JACV-CLS instances and 240 JACV-ADV instances. The 80 anchors that did not yield complete pairs were excluded because their counterfactuals failed all admissibility criteria across the fallback chain (trivial similarity, excessive lexical overlap, or no legally significant change despite multiple generation attempts). Table 2 shows the corpus statistics; the strong length heterogeneity across fields, particularly in law and relief, was an important factor in benchmark formulation.

Preprocessing employed three-level anonymization combining regular patterns for personal and procedural identifiers, specialized lists of sensitive domain terms, and automatic entity recognition to mask persons, organizations and locations. Data handling followed LGPD principles and CNJ AI governance guidelines [Brasil 2018, Conselho Nacional de Justiça 2025b].

3.2. JACV Benchmark Construction

Each of the 200 sampled anchors was used to generate two controlled counterfactual reliefs: one of type `INCOHERENT` (a relief that does not follow from the legal reasoning but does not directly contradict it) and one of type `CONTRADICTORY` (a relief that directly inverts the legal reasoning). Generation used the fact, law, and original relief fields as context but restricted modifications to the relief field. Sabiá-4 [Laitz et al. 2026] served as the primary generator, Claude-Sonnet-4.6 [Anthropic 2026b] as the fallback, and Claude-Opus-4.7 [Anthropic 2026a] as the cross-validation model (all accessed via API in March–April 2026). Outputs were rejected when they reproduced the original relief, preserved excessive lexical overlap, or failed to introduce a legally significant change.

Table 3. Illustrative example of legal reasoning (Law) and relational relief counterfactuals in JACV.

Field / Label	Text Content (Brazilian Portuguese Excerpt)
Law (Excerpt)	*"...A busca domiciliar realizada sem mandado judicial e respaldada tão somente em denúncia anônima viola o Art. 5º, XI da Constituição Federal, tornando ilícitas todas as provas dela derivadas e impondo a absolvição do réu..."*
COHERENT	Declaração de ilicitude das provas obtidas na busca domiciliar e consequente absolvição do réu por ausência de prova da materialidade.
INCOHERENT	Condenação do réu ao pagamento de indenização por danos morais no valor de R$ 50.000,00 acrescidos de juros e correção monetária.
CONTRADICTORY	Manutenção da validade das provas colhidas na busca domiciliar e subsistência da condenação criminal do réu em todos os seus termos.

Each accepted counterfactual was further validated by a model distinct from the generator as an admissibility filter.

The benchmark comprises two complementary tasks. In **JACV-CLS**, the model classifies a (law, relief) pair as `COHERENT`, `INCOHERENT` or `CONTRADICTORY`. In **JACV-ADV**, the model selects the coherent relief from two alternatives for a given law field:

$$\hat{y}_{adv} = \begin{cases} A, & \text{if } s(d, p_A) \geq s(d, p_B), \\ B, & \text{otherwise,} \end{cases}$$

where $s(d, p)$ denotes the coherence score for the law–relief pair d, p. Anchor sampling was balanced between appellate and non-appellate, and the position of the correct alternative in JACV-ADV was randomized. A key assumption is that original documents represent coherent law–relief pairs, partially supported by their institutional context (court-accepted documents satisfy a minimum formal adequacy standard) and by the pilot expert validation reported in Section 4.3.

To motivate the necessity of legal-argumentative coherence verification over surface textual proximity, Table 3 presents an illustrative benchmark anchor along with its three relational relief variants. In this example, the legal reasoning establishes that a warrantless home search based solely on an anonymous tip is unconstitutional under Art. 5, XI of the Brazilian Federal Constitution. A model evaluating surface lexical overlap alone might assign high similarity scores to both the `COHERENT` and `CONTRADICTORY` reliefs because both extensively share domain terminology (*e.g.*, "busca domiciliar", "provas", "condenação"). However, while the `COHERENT` relief logically requests evidence suppression and acquittal, the `CONTRADICTORY` relief requests the opposite outcome (upholding the conviction). Conversely, the `INCOHERENT` relief introduces an off-topic demand for civil compensation (moral damages), shifting the domain scope. Detecting these distinctions requires models to track the inferential direction between legal premises and relief rather than relying on surface similarity.

3.3. Evaluated Models

The experiments compared five families of methods. **Statistical baselines**: random classifier, majority class, and TF-IDF with logistic regression (TF-IDF-LogReg), used

to estimate resolvability through surface regularities. **Feature extractor encoders** (frozen transformer weights, trained linear head): GAIA [Camilo-Junior et al. 2025], JurisBERT [Viegas et al. 2023], BERTimbau-base [Souza et al. 2020], and LegalBERT-pt [Silveira et al. 2023]. **Fine-tuned encoders** (end-to-end fine-tuning on the benchmark): RoBERTaLexPT-base [Garcia et al. 2024] and multilingual BERT [Devlin et al. 2018], both using the concatenated [law; `[SEP]`; relief] input for sequence classification. **LLM zero-shot**: Sabiá-4 [Laitz et al. 2026], Claude-Sonnet-4.6 [Anthropic 2026b], Claude-Opus-4.7 [Anthropic 2026a], and Mistral-Large-3 [Mistral AI 2025]. **Legal-CoT**: structured prompting decomposing reasoning into relevant facts, applicable norm, logical connection, and final conclusion, applied to Sabiá-4, Claude-Sonnet-4.6, and Mistral-Large-3.

Model selection was guided by three criteria: availability and relevance for pt-BR (BERTimbau, LegalBERT-pt, JurisBERT, RoBERTaLexPT, Sabiá-4), cross-lingual coverage (multilingual BERT, GAIA, Mistral-Large-3), and frontier capability for upper-bound estimation (Claude-Sonnet-4.6, Claude-Opus-4.7). Fine-tuning used three epochs, learning rate 2×10^{-5}, weight decay 0.01, maximum 512 tokens, and gradient accumulation on an NVIDIA H100 GPU. For LLMs, each JACV-ADV instance was presented as a single prompt with both alternatives; for encoder-based models, each (law, relief) pair was scored independently using the predicted `COHERENT` probability, with the higher-scoring alternative selected.

3.4. Experimental Protocol and Statistical Analysis

Supervised evaluations used grouped cross-validation with three partitions, preserving the source document identifier as the grouping unit to prevent leakage (multiple instances may derive from the same anchor). Main metrics were accuracy and macro-F1 for JACV-CLS, and selection accuracy for JACV-ADV. All accuracy values were computed over the full instance set (parsing errors counted as incorrect); conditional accuracy over answered-only instances is reported separately. Confidence intervals were obtained by non-parametric bootstrap with 2,000 resamples. The protocol also recorded the **parsing error rate** (proportion of out-of-format responses) and **abstention rate**, which are relevant in legal settings where format failures can be as consequential as classification errors.

3.5. Code and Data Availability

All experimental code, evaluation scripts, prompts, and benchmark instances are available in the public repository.[1] The original procedural corpus is not publicly released because it derives from authentic judicial documents and remains subject to privacy, confidentiality, and LGPD constraints. Researchers interested in academic use of the corpus may contact the authors to verify availability under a formal data-sharing agreement, subject to institutional authorization and compliance with Brazilian data protection law.

4. Results and Discussion

4.1. Global Performance and Method Families

Table 4 summarizes performance across all five families. Among LLMs in zero-shot mode, **Claude-Opus-4.7** achieved the best adversarial result at **99.58%** on JACV-ADV,

[1] `https://github.com/Willgnner-Santos/JACV`

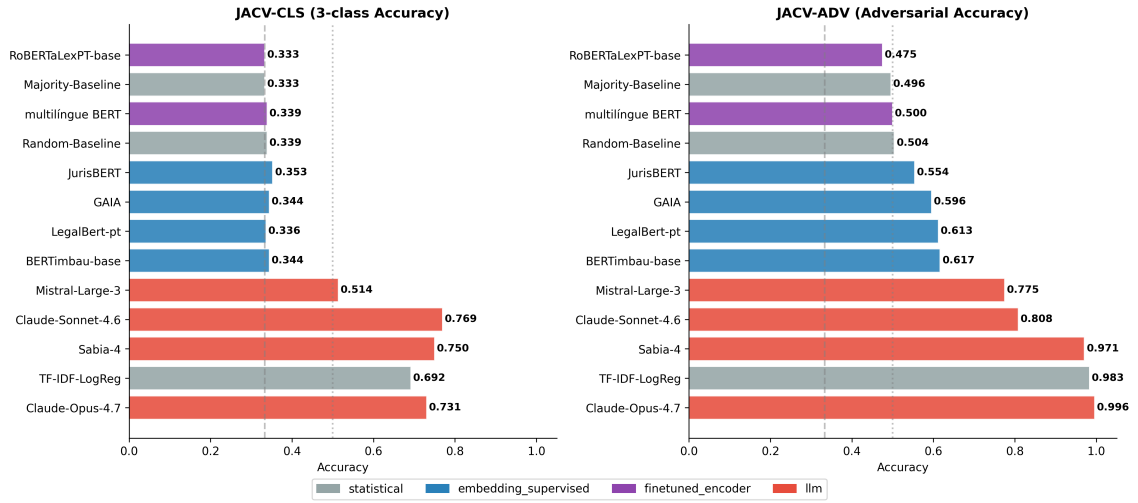


Figure 1. Global comparison between JACV-CLS and JACV-ADV accuracy across model families.

and **Claude-Sonnet-4.6** the best on JACV-CLS at **76.94%** accuracy (0.766 macro-F1), followed closely by **Sabiá-4** at 75.00% and 0.747, respectively. Figures 1 and 2 present the comparative performance across models ordered by performance and family. Embedding-based and fine-tuned models remain near chance level on JACV-CLS (33–35%), while the best LLMs show substantial gains. Supervised fine-tuning did not improve over frozen feature extractor encoders. Beyond dataset size, input truncation represents a major structural constraint: standard encoder baselines were evaluated under a 512-token sequence limit, whereas the 95th-percentile length for the law field in our corpus reaches 11,099 tokens (Table 2). Consequently, encoders process only the head of the legal argumentation, discarding crucial inferential premises. Evaluating encoders under long-context architectures (*e.g.*, Longformer or extended-window Transformers) is necessary to isolate inherent architectural capacity from sequence length limitations.

TF-IDF-LogReg achieved **98.33%** on JACV-ADV and **69.2%** on JACV-CLS, outperforming all encoder families and surpassing Claude-Sonnet-4.6 and Mistral-Large-3 in raw adversarial accuracy. This indicates that a non-trivial portion of JACV-ADV remains solvable through surface cues introduced by the generation process. JACV should not be read as isolated evidence of deep legal reasoning; rather, it simultaneously captures argumentative-coherence signals and residual lexical separability, whose interaction must be considered when interpreting the findings. That said, the three-class task is harder for the lexical baseline (69.2% vs. 76.9% for Claude-Sonnet-4.6), indicating that JACV-CLS requires finer semantic discrimination not reducible to lexical patterns.

4.2. Operational Robustness and Legal CoT

For **Claude-Sonnet-4.6** and **Mistral-Large-3** in zero-shot mode, the main limitation was not inferential but one of format adherence: their raw ADV accuracies (80.83% and 77.50%) are substantially below their conditional accuracies over valid responses (99.49% and 97.38%, respectively). In these cases, the bottleneck was the response protocol rather than the underlying coherence judgment.

Legal CoT produced a markedly worse outcome. Raw accuracies for the three CoT models ranged from 20.42% (Mistral-Large-3-CoT) to 27.08% (Sabiá-4-CoT), with

Table 4. Main results on JACV-CLS and JACV-ADV. ADV Err = proportion of out-of-format responses; ADV (ans.) = accuracy over valid responses only. Bold: best per metric. 95% CIs from bootstrap (2,000 resamples); parsing errors counted as incorrect in all accuracy figures.

Family	Model	CLS Acc. [95% CI]	CLS F1 [95% CI]	ADV Acc. [95% CI]	ADV Err (%)	ADV (ans.)
Statistical	Random	0.339 [0.292, 0.389]	0.339 [0.290, 0.389]	0.504 [0.442, 0.567]	–	–
	Majority	0.333 [0.283, 0.383]	0.167 [0.147, 0.185]	0.496 [0.433, 0.558]	–	–
	TF-IDF-LogReg	0.692 [0.647, 0.739]	0.683 [0.636, 0.729]	0.983 [0.963, 0.996]	0.0	–
Feat. extr.	GAIA	0.344 [0.297, 0.394]	0.342 [0.293, 0.390]	0.596 [0.533, 0.654]	0.0	–
	JurisBERT	0.353 [0.308, 0.406]	0.351 [0.304, 0.401]	0.554 [0.488, 0.617]	0.0	–
	BERTimbau	0.344 [0.297, 0.394]	0.333 [0.286, 0.381]	0.617 [0.554, 0.679]	0.0	–
	LegalBERT-pt	0.336 [0.289, 0.383]	0.334 [0.286, 0.382]	0.613 [0.550, 0.675]	0.0	–
Fine-tuning	RoBERTaLexPT	0.333 [0.289, 0.383]	0.333 [0.289, 0.382]	0.475 [0.408, 0.538]	0.0	–
	mBERT	0.339 [0.292, 0.386]	0.276 [0.236, 0.316]	0.500 [0.438, 0.563]	0.0	–
LLM zero-shot	Sabiá-4	0.750 [0.703, 0.792]	0.747 [0.699, 0.787]	0.971 [0.946, 0.992]	0.0	–
	Claude-S-4.6	**0.769** [0.725, 0.811]	**0.766** [0.723, 0.805]	0.808 [0.758, 0.854]	18.8	0.995
	Claude-O-4.7	0.731 [0.683, 0.772]	0.721 [0.675, 0.760]	**0.996** [0.988, 1.000]	0.0	–
	Mistral-L-3	0.514 [0.458, 0.567]	0.438 [0.398, 0.474]	0.775 [0.721, 0.825]	20.4	0.974
Legal-CoT	Sabiá-4-CoT†	–	–	0.271 [0.213, 0.333]	47.9	0.520
	Sabiá-4-CoT fixed‡	–	–	0.304 [0.246, 0.363]	54.2	0.664
	Claude-S-4.6-CoT	–	–	0.250 [0.196, 0.304]	47.5	0.476
	Mistral-L-3-CoT	–	–	0.204 [0.154, 0.254]	58.3	0.490

† Affected by output-length bug (`max_tokens=20`); see Section 4.2.
‡ Corrected re-run (`max_tokens=400`); parsing error rate did not decrease, confirming a protocol-level limitation.

parsing error rates between 47.5% and 58.3%. None used the `INCONCLUSIVE` option explicitly permitted in the prompt. A post-hoc review revealed that the Sabiá-4-CoT run was additionally affected by an API configuration error: `max_tokens=20`, far below what is needed for a structured reasoning chain. A corrected run (`max_tokens=400`) yielded raw ADV accuracy of **30.42%** [95% CI: 0.246, 0.363], parsing error rate of **54.2%**, and conditional accuracy of **66.4%**. The conditional accuracy improved substantially (52.0% $\rightarrow$ 66.4%), confirming that the token constraint did degrade response quality; however, the parsing error rate did not decrease; it increased slightly (47.9% $\rightarrow$ 54.2%), indicating that the format-adherence problem is intrinsic to the multi-block CoT prompt design, not caused by truncation. The Claude-Sonnet-4.6-CoT and Mistral-Large-3-CoT results, unaffected by the token constraint, show similar error rates (47.5% and 58.3%),

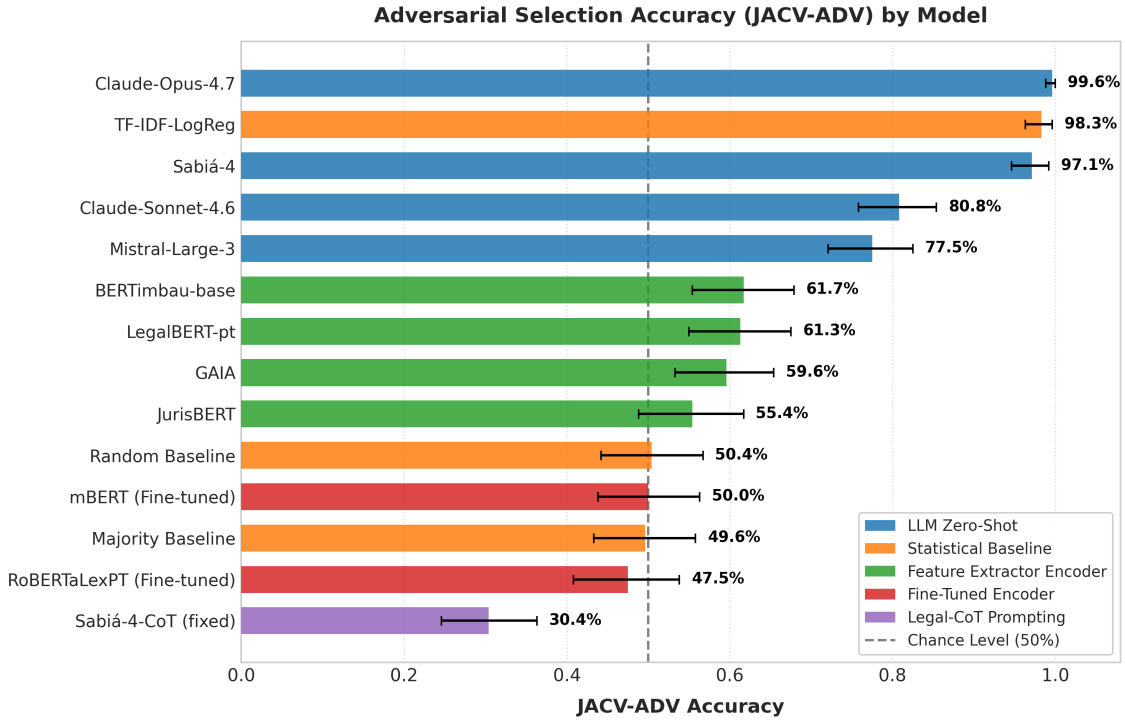


Figure 2. Adversarial selection accuracy (JACV-ADV) across evaluated models, ordered by performance and categorized by method family, with 95% bootstrap confidence intervals.

further confirming this is a protocol-level limitation.

These results show that Legal-CoT, in the design employed here (requiring discursive blocks for facts, applicable norm, logical connection, and conclusion), substantially increases vulnerability to poorly formatted responses without producing a compensating gain in accuracy. This does not invalidate structured legal prompting in general, but demonstrates that instruction calibration and strict output control are prerequisites for its operational effectiveness.

4.3. Analysis by Class, Perturbation Type, and Expert Validation

Figure 3 shows that model behavior on JACV-CLS was not homogeneous across classes. Among the best LLMs, `COHERENT` was the most stable class and `INCOHERENT` the most difficult. A closer examination of Claude-Sonnet-4.6 errors reveals a characteristic pattern: nearly all `INCOHERENT` errors consist of the model predicting `CONTRADICTORY` (57 of 57 `INCOHERENT` errors), while `COHERENT` and `CONTRADICTORY` errors are more evenly distributed. The `INCOHERENT`/`CONTRADICTORY` boundary is the hardest distinction in JACV-CLS and is consistent across all LLMs (cf. Figure 3). A representative error involved an appellate document discussing the validity of a warrantless home entry based on an anonymous tip (legal reasoning: such entry is invalid absent concrete indicators of a flagrant offense); the model predicted the counterfactual relief as `COHERENT` when the label was `CONTRADICTORY`: the perturbation requested confirmation of the search's validity, directly inverting the legal reasoning, a conflict that required recognizing directional opposition rather than mere mismatch.

TF-IDF feature importance analysis further illuminates the lexical separability in JACV-ADV (Figure 4). The most discriminative features for `INCOHERENT` are

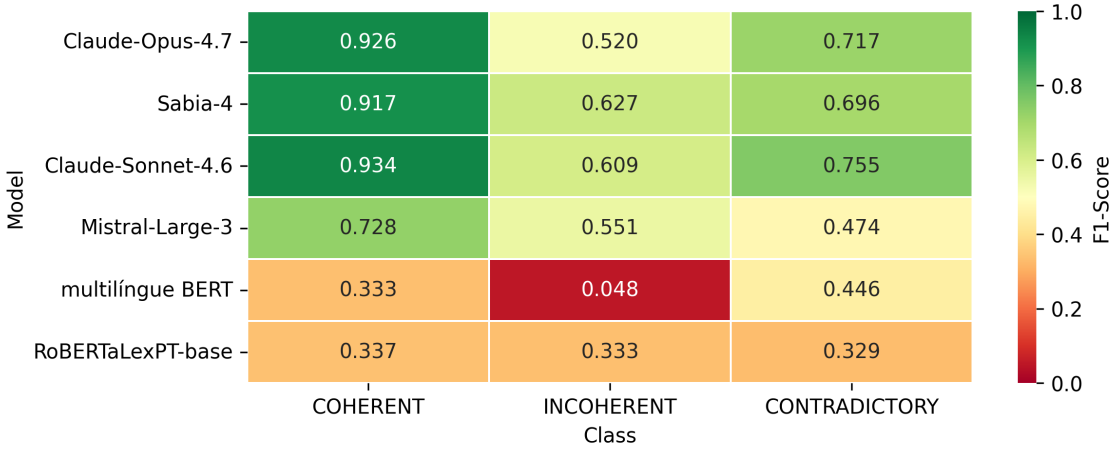


Figure 3. F1 per class for LLMs and fine-tuned encoders on JACV-CLS.

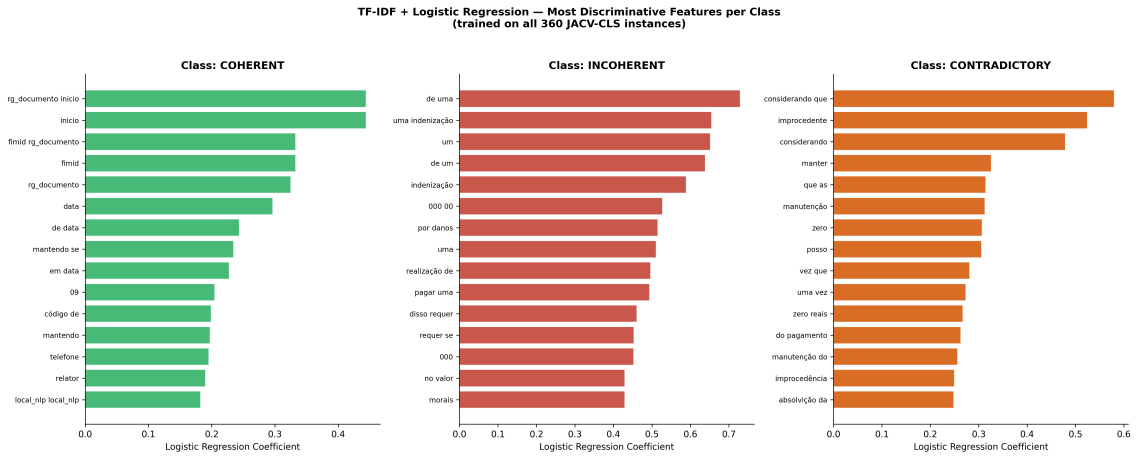


Figure 4. Top-15 most discriminative TF-IDF features per class (logistic regression coefficients). `INCOHERENT`: compensation-related terms; `CONTRADICTORY`: outcome-negating terms; `COHERENT`: document-closing markers and anonymization artifacts from original reliefs.

compensation-related terms (“indenização”, “por danos”, “morais”, “pagar uma”), reflecting the strategy of inserting mismatched compensation requests. `CONTRADICTORY` is dominated by outcome-negating terms (“improcedente”, “absolvição da”, “considerando que”), consistent with perturbations that invert the requested outcome. `COHERENT` features anonymization artifacts (“rg_documento”, “fimid”) and document-closing markers (“mantendo”, “relator”) present in the original reliefs. These patterns confirm that a non-trivial portion of adversarial separability stems from lexical artifacts introduced by the generation process, rather than purely from legal-argumentative reasoning. To address this in future benchmark revisions, counterfactual generation protocols can be constrained to preserve surface vocabulary while explicitly altering argumentative directionality or party polarity.

As an exploratory construct validity check, 10 benchmark instances (4 JACV-CLS and 6 JACV-ADV) were independently evaluated by three legal experts, covering all classes, both perturbation types and both document types. Experts agreed with benchmark labels in 9 of 10 instances (90% agreement). The single disputed instance

involved a CONTRADICTORY label where two experts assigned INCOHERENT, a boundary case involving partial relief overlap, consistent with the model-level difficulty at the INCOHERENT/CONTRADICTORY boundary reported above. Inter-annotator agreement among the three experts reached $\kappa = 0.87$ (Fleiss' κ), indicating strong professional consistency. While this pilot provides valuable preliminary empirical support that benchmark contrasts are recognizable under human professional judgment, it does not substitute for a full-scale annotation campaign. Expanding human validation across diverse jurisdictions and legal domains represents a key direction for future work.

5. Conclusion

This paper presented JACV, a benchmark and evaluation protocol for legal-argumentative coherence in Brazilian Portuguese procedural documents. Its central motivation is that strong performance in legal classification, retrieval, or thematic labeling does not necessarily imply the ability to verify whether a requested relief is coherent with the legal reasoning of a document. By constructing controlled counterfactual examples from authentic procedural documents, JACV shifts the evaluation focus from textual proximity to the compatibility between legal reasoning and relief.

The results show that legal-argumentative coherence is a relevant and empirically observable dimension for evaluating AI in judicial contexts. The best LLMs achieved strong performance, with Claude-Opus-4.7 reaching 99.58% on JACV-ADV and Claude-Sonnet-4.6 reaching 76.94% on JACV-CLS, substantially outperforming encoder-based approaches. However, the strong TF-IDF baseline indicates that part of the benchmark remains lexically separable, cautioning against interpreting high adversarial accuracy as definitive evidence of deep legal reasoning. Encoder models were also limited by the 512-token sequence constraint relative to long legal reasoning fields ($P_{95} = 11,099$ tokens), reinforcing the need for long-context architectures. In addition, Legal-CoT degraded operational robustness, with parsing error rates above 47% across all models; a corrected Sabiá-4-CoT run confirmed that this limitation stems from the multi-block prompt design rather than token truncation alone.

JACV therefore provides an initial framework for evaluating legal-argumentative coherence in Brazilian Portuguese, combining authentic procedural data, controlled counterfactual construction, and comparison across five method families. Future work should: (i) expand the corpus across State and Federal courts and legal branches; (ii) conduct large-scale human annotation with multi-annotator agreement; (iii) design constrained perturbations that preserve surface vocabulary to better isolate logical reasoning; (iv) evaluate long-context encoder architectures; and (v) investigate calibrated abstention and output-constrained prompting strategies.

6. Acknowledgments

This work was supported by Centro de Excelência em Inteligência Artificial. This work was also supported by the National Institute of Science and Technology (INCT) in Responsible Artificial Intelligence for Computational Linguistics and Information Treatment and Dissemination (TILD-IAR) grant number 408490/2024-1.

References

Anthropic (2026a). Claude Opus 4.7. Access em: 24 abr. 2026.

Anthropic (2026b). System Card: Claude Sonnet 4.6. Access em: 24 abr. 2026.

Brasil (2018). Lei nº 13.709, de 14 de agosto de 2018. lei geral de proteção de dados pessoais (lgpd). https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm. Access em 22 abr. 2026.

Brito, M., Pinheiro, V., Furtado, V., Monteiro Neto, J., Bomfim, F., Costa, A. d., and Silveira, R. (2023). CDJUR-BR - uma coleção dourada do judiciário brasileiro com entidades nomeadas refinadas. In *Proceedings of the 14th Brazilian Symposium in Information and Human Language Technology*, pages 176–195, Belo Horizonte, Brasil. Association for Computational Linguistics.

Camilo-Junior, C. G., Oliveira, S. S. T., Pereira, L. A., Amadeus, M., Scotti, R., Fazzioni, D., Novais, A. M. A., and Jordão, S. A. A. (2025). GAIA: An open language model for brazilian portuguese. https://huggingface.co/CEIA-UFG/Gemma-3-Gaia-PT-BR-4b-it.

Chu, Y. Y., Huang, S.-c., and Shao, H.-L. (2025). Unpacking legal reasoning in LLMs: Chain-of-thought as a key to human-machine alignment in essay-based NLU tasks. In *Proceedings of the 5th Workshop on Natural Logic Meets Machine Learning (NALOMA)*, pages 1–7, Bochum, Alemanha. Association for Computational Linguistics.

Conselho Nacional de Justiça (2025a). Justiça em números: Judiciário reduziu acervo e alcançou produtividade histórica em 2024. https://www.cnj.jus.br/justica-em-numeros-judiciario-reduziu-acervo-e-alcancou-produtividade-historica-em-2024/. Access em 22 abr. 2026.

Conselho Nacional de Justiça (2025b). Resolução nº 615, de 11 de março de 2025. https://atos.cnj.jus.br/files/original1555302025031467d4517244566.pdf. Access em 22 abr. 2026.

Dahl, M., Magesh, V., Suzgun, M., and Ho, D. E. (2024). Large legal fictions: Profiling legal hallucinations in large language models. *arXiv preprint arXiv:2401.01301*.

Devlin, J., Chang, M., Lee, K., and Toutanova, K. (2018). BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Fei, Z., Shen, X., Zhu, D., Zhou, F., Han, Z., Huang, A., Zhang, S., Chen, K., Yin, Z., Shen, Z., Ge, J., and Ng, V. (2024). Lawbench: Benchmarking legal knowledge of large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Miami, Florida, USA. Association for Computational Linguistics.

Garcia, E. A. S., Silva, N. F. F., Siqueira, F., Albuquerque, H. O., Gomes, J. R. S., Souza, E., and Lima, E. A. (2024). RoBERTaLexPT: A legal RoBERTa model pretrained with deduplication for Portuguese. In Gamallo, P., Claro, D., Teixeira, A., Real, L., Garcia, M., Oliveira, H. G., and Amaro, R., editors, *Proceedings of the 16th International Conference on Computational Processing of Portuguese*, pages 374–383, Santiago de Compostela, Galicia/Spain. Association for Computational Lingustics.

Goebel, R., Kano, Y., Kim, M.-Y., Rabelo, J., Satoh, K., and Yoshioka, M. (2024). Overview of benchmark datasets and methods for the legal information extraction/entailment competition (COLIEE) 2024. In *JSAI International Symposium on Artificial Intelligence*, pages 109–124. Springer.

Guha, N., Nyarko, J., Ho, D. E., Ré, C., Chilton, A., Narayana, A., Chohlas-Wood, A., Peters, A., Waldon, B., Rockmore, D. N., et al. (2023). Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 36.

Laitz, T., Almeida, T. S., Abonizio, H., Junior, R. M., Bonás, G. K., Piau, M., Larcher, C., Pires, R., and Nogueira, R. (2026). Sabi\'a-4 technical report. *arXiv preprint arXiv:2603.10213*.

Luz de Araujo, P. H., de Campos, T. E., Ataides Braz, F., and Correia da Silva, N. (2020). VICTOR: a dataset for brazilian legal documents classification. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1449–1458, Marseille, France. European Language Resources Association.

McCoy, R. T., Pavlick, E., and Linzen, T. (2019). Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448. Association for Computational Linguistics.

Mistral AI (2025). Mistral Large 3. Access em: 24 abr. 2026.

Polo, F. M., Mendonça, G. C. F., Parreira, K. C. J., Gianvechio, L. d. G., Cordeiro, P. S., Ferreira, J. B., Lima, L. M. P. d., Maia, A. C. d. A., and Vicente, R. (2021). Legalnlp: Natural language processing methods for the brazilian legal language. In *Anais do Encontro Nacional de Inteligência Artificial e Computacional*, Porto Alegre, Brasil. SBC.

Servantez, S., Barrow, J., Hammond, K., and Jain, R. (2024). Chain of logic: Rule-based reasoning with large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2721–2733, Bangkok, Tailândia. Association for Computational Linguistics.

Silveira, R., Ponte, C., Almeida, V., Pinheiro, V., and Furtado, V. (2023). Legalbert-pt: A pretrained language model for the brazilian portuguese legal domain. In *Brazilian Conference on Intelligent Systems*, pages 268–282. Springer.

Souza, F., Nogueira, R., and Lotufo, R. (2020). Bertimbau: Pretrained bert models for brazilian portuguese. In *Intelligent Systems*, volume 12319 of *Lecture Notes in Computer Science*, pages 403–417. Springer.

T.y.s.s., S., Haddad, R., and Grabmair, M. (2024). ECtHR-PCR: A dataset for precedent understanding and prior case retrieval in the european court of human rights. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5473–5483, Torino, Italia. ELRA and ICCL.

Viegas, C. F. O., Costa, B. C., and Ishii, R. P. (2023). JurisBERT: A new approach that converts a classification corpus into an STS one. In *Computational Science and Its Applications – ICCSA 2023*, pages 349–365. Springer Nature Switzerland.

Yu, F., Quartey, L., and Schilder, F. (2023). Exploring the effectiveness of prompt engineering for legal reasoning tasks. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13582–13596, Toronto, Canada. Association for Computational Linguistics.