Fact, Legal Thesis, and Noise: Comparing Classical ML, BERT, and LLMs for Triage of Initial Petitions in Brazilian Portuguese

Willgnner F. Santos[1,2], João G. G. Viana[2], Adriel Mori[1], Manuel L. Zengo[1,2,3], Milton P. de Ávila Júnior[2], Antônio P. de Castro Júnior[2], Fernando R. Trindade[2], Valdemar V. G. Neto[1] and Ronaldo M. da Costa[1]

[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

{willgnner_santos,adrielmori}@discente.ufg.br

{jggviana202,sir.lucala,miltonavila.dev}@gmail.com

{apcastrojr,vgracianoneto,ronaldomc12}@gmail.com

Abstract. *In Brazilian state courts, manual document triage is slow and error-prone, requiring the separation of factual narrative, legal arguments, and procedural noise. This work compares classical models, fine-tuned BERT encoders, and out-of-the-box large language models (LLMs) for classifying Brazilian Portuguese initial petitions into fact, legal thesis, and noise. To our knowledge, this is the first such comparison for this task in Portuguese. Portuguese-pretrained encoders performed best (macro-F1 about* 0.99*), followed by a long-context Longformer (*0.983*) and classical models (*0.973*); out-of-the-box LLMs reached about* 0.85 *and were highly sensitive to output format, with formatting failures significantly impacting performance. Errors concentrate on passages mixing factual narration and legal argument, confirmed by three legal professionals (Fleiss'* $\kappa = 0.46$*). An ablation shows that anonymization markers do not drive the scores. The results help guide model selection for court document triage.*

1. Introduction

Document triage is one of the first steps in court operations and conditions all subsequent analysis [Santos et al. 2026]. At the Court of Justice of the State of Goiás (TJGO) that motivates this study, approximately ten thousand documents arrive each day, making manual classification slow, error-prone, and subject to inconsistent criteria across staff, while increasing waiting times for people seeking judicial assistance. The problem is compounded by the unstructured nature of legal texts, which interleave factual descriptions, legal arguments, and content of no analytical value. Automatically separating fact, legal thesis, and noise can support triage and reduce manual effort without removing human responsibility [Kalamkar et al. 2022, Aragy et al. 2021].

Natural language processing (NLP) and machine learning have been applied to legal documents in tasks such as text segmentation, rhetorical role identification, and sentence classification, with models ranging from support vector machines

to Transformer-based architectures [Šavelka and Ashley 2018, Bhattacharya et al. 2023]. For Portuguese, BERT models pretrained on the language, such as BERTimbau, provided more suitable representations [Souza et al. 2020], and more recently large language models (LLMs) have been evaluated on legal tasks in Brazilian courts [Dal Pont et al. 2026, Bomfim et al. 2026].

Despite this progress, no study has directly compared these three families for classifying Brazilian Portuguese legal text into fact, legal thesis, and noise. Most work focuses on a single family, adopts different label schemes, or addresses English-speaking jurisdictions. The closest work [Aragy et al. 2021] examined rhetorical roles in Portuguese petitions using a different schema and a single model family on a much smaller corpus. To our knowledge, this is the first comparison of classical models, fine-tuned encoders, and LLMs on the same petition corpus under this schema, grounded in Brazilian civil procedure. The trade-off among performance, computational cost, and ease of deployment has also received little attention. Our research question is: to what extent do different AI approaches classify Brazilian Portuguese legal text into fact, legal thesis, and noise, considering both performance and cost?

We compare classical models, BERT-based models, and LLMs on this task using the same dataset, a design also adopted for legal documents in a related setting [Santos et al. 2025]. The study does not evaluate production deployment and does not replace human analysis; its focus is the comparison itself. The experiments show differences in performance and cost that help guide model selection in resource-constrained settings, relevant to both the legal NLP community and institutions processing large document volumes.

2. Related Work

NLP and machine learning applied to legal texts cover classification, segmentation, and document retrieval, with particular emphasis on rhetorical role identification, that is, assigning functions such as fact, argument, and decision to sentences or passages in a document [Lou et al. 2021, Adsul and Deshmukh 2023]. This is close to our problem, namely separating factual passages from legal theses and ancillary content in loosely structured documents, which supports downstream tasks such as summarization and case organization [Darji et al. 2025].

On the technical side, the literature has evolved from classical models [Mok et al. 2021] to Transformer-based architectures [Devlin et al. 2019]. Studies indicate that contextual representations, in particular BERT-based models, improve legal sentence classification and cross-domain generalization, though gains depend on the annotation scheme [Csányi et al. 2025, Malik et al. 2022]. Benefits have also been reported from incorporating sentence position into BERT models [Belfathi et al. 2023] and from deep semantic segmentation for passage labeling [Mohan and Nair 2024]. Because legal documents tend to be long, the 512-token limit of standard BERT motivates the use of long-context encoders such as Longformer [Beltagy et al. 2020] and BigBird [Zaheer et al. 2020], as well as hierarchical approaches, both of which are evaluated or discussed in this paper.

More recently, LLMs have been applied to legal tasks including fact extraction, legal reasoning, and agent-based systems [Bendová et al. 2025, Bonfim et al. 2026,

Tomasi 2026, Liepin et al. 2025], with prompt engineering surveys covering evidence-constrained and decomposed strategies that we also adopt, while noting that retrieval-augmented and chain-of-thought variants remain underexplored [Lai et al. 2024]. LegalBench.PT [Canaverde et al. 2025] targets doctrinal knowledge in European Portuguese, not petition triage, making our focus complementary; the Cambridge Law Corpus [Östling et al. 2023] illustrates the value of restricted-access legal corpora for privacy, an approach we follow. The present study brings all three families onto a single Brazilian petition corpus with explicit attention to performance and cost.

3. Methodology

Figure 1 summarizes the pipeline, from data preparation to the comparison of the three machine learning families (classical ML, fine-tuned BERT encoders, and LLMs).

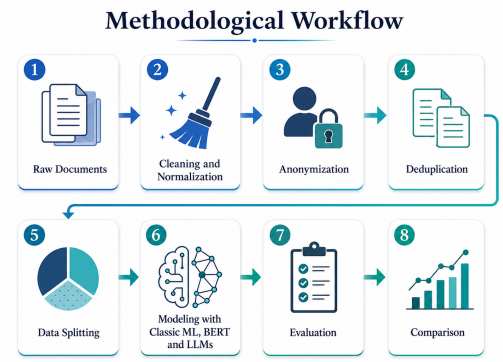


Figure 1. Overview of the methodological pipeline adopted in the study.

Study design and definitions. This study compares classical ML, fine-tuned BERT encoders, and LLMs on the same task. A document is an initial petition; a passage (or chunk) is the textual unit extracted by the institutional system at TJGO. Each passage receives a single label, and metrics are computed at the passage level. The classes are: fact (events, circumstances, concrete case elements), legal thesis (arguments, statutory basis, requests), and noise (headers, procedural data, boilerplate). Data was split (train/validation/test) stratified by class, grouping passages from the same document in the same split to avoid leakage (seed $= 42$). The test set was held out during fine-tuning. Vectorizers were fitted on training data only. Validation sets were used strictly for early stopping and hyperparameter selection.

Corpus. The study is based on an unpublished corpus of initial petitions in Brazilian Portuguese, provided by the Court of Justice of the State of Goiás (TJGO) and processed through a cleaning, normalization, deduplication, and anonymization pipeline. Labels were produced under explicit institutional guidelines and reviewed by domain experts at TJGO. The raw data contained $26,781$ records; after removing empty records and exact duplicates ($1,298$ occurrences), the final corpus comprises $25,482$ labeled passages across three naturally balanced classes, with a slight majority of noise (Table 1). No artificial class balancing was performed during preprocessing. Subject to institutional authorization and a confidentiality agreement, the anonymized dataset (placeholder tokens and class labels, no original sensitive values) may be released to identified researchers for

academic use, in line with LGPD [Rapôso et al. 2019] and CNJ Resolution No. 615/2025 [Conselho Nacional de Justiça (Brasil) 2025].

Table 1. Class distribution and basic length statistics (in words) in the corpus after cleaning and deduplication.

Class	Count	%	Min	Max	Mean	Median
Fact	8,517	33.4%	1	14,021	493.1	352
Legal thesis	8,039	31.5%	1	12,874	610.4	481
Noise	8,926	35.0%	1	45,786	1,698.2	1,424
Total (Passages)	25,482	100%	1	45,786	952.3	687
Total (Docs)	4,792	–	2	45,786	1,450.0	815

Measured in tokens with the BERTimbau-base tokenizer on a sample of $3,000$ texts, the passage median is $1,335$ and about 79.4% exceed the 512-token BERT limit. Length varies by class, with fact passages the shortest and noise the longest, a difference that is informative but calls for caution, since length may itself act as a superficial cue.

Preprocessing, anonymization, and leakage risk. The same preprocessing pipeline was applied before training or evaluating any model: reading raw files; removing empty records and exact duplicates; normalizing encoding, whitespace, and line breaks; and validating labels. The classical, BERT, and LLM pipelines reuse these routines. The class mapping is $0 =$ legal thesis, $1 =$ fact, $2 =$ noise. Anonymization reduces the risk of personal data exposure. It uses regular expressions to mask sensitive entities with standardized placeholders across nine categories: CPF, CNPJ, email, phone, postal code, OAB registration number, monetary values, URLs, and case numbers. In the full corpus, $463,535$ occurrences were masked, averaging approximately 17 entities per document, and the audit log records only aggregate counts by type, without storing the original values.

Placeholders are not evenly distributed: in the training split, noise passages average 26.9 tags (median 17; 98.7% have at least three), against 15.4 for legal thesis and 7.8 for fact. This reflects structural differences, since noise passages are mostly procedural headers. The potential impact of this lexical shortcut is evaluated in Section 4.

Deduplication removed only exact matches. To gauge the near-duplicate risk, we compared each test passage against a random sample of $2,000$ training passages using 5-gram Jaccard similarity. About 19.7% of test passages had a match above $J = 0.5$ and 10.9% above $J = 0.8$. Initial petitions are formulaic, with boilerplate requests and statutory references recurring across unrelated cases, so much of this similarity comes from shared boilerplate rather than copied content. Grouping by document keeps passages from the same petition in a single split, and shared boilerplate tends to carry the same label, usually noise. Even so, these rates are a limitation. Since this limited audit lacked a per-class breakdown, score inflation for the fact and legal-thesis classes cannot be fully excluded. Labels annotated under explicit institutional criteria may produce predictable boundaries, helping explain why even bag-of-words models do well.

Classical models. The classical approach uses TF-IDF representations over words, with n-grams from 1 to 2, a minimum document frequency of 2, and a vocabulary of up to $50,000$ terms; accents were preserved. Eight classifiers were evaluated: Logistic Regres-

sion, linear SVM (`LinearSVC`), RBF-kernel SVM, Multinomial Naive Bayes, Random Forest, Extra Trees, LightGBM, and XGBoost. Probabilities for the linear SVM were obtained by calibration. Hyperparameters were set based on established defaults for text classification in Portuguese without systematic grid search. The best classical model was selected by stratified five-fold cross-validation on the training set using macro-F1.

BERT-based models. Eight encoder models were fine-tuned: BERTimbau-base and BERTimbau-large; RoBERTaLexPT-base and RoBERTaCrawlPT-base [Garcia et al. 2024]; XLM-RoBERTa-base [Conneau et al. 2019]; mBERT [Devlin et al. 2019]; mDeBERTa-v3-base [He et al. 2021a, He et al. 2021b]; and Longformer-base-4096 [Beltagy et al. 2020]. The first three are pretrained on Portuguese, with RoBERTaLexPT targeting the legal domain. Fine-tuning used each model's tokenizer, dynamic padding, AdamW (weight decay 0.01), a linear scheduler with warmup, cross-entropy loss, and early stopping (patience 2, on macro-F1). The seven standard models used a maximum length of 512 tokens with simple truncation and up to 10 epochs. Longformer used $4,096$ tokens, global attention on the `[CLS]` token for sequence classification, and 5 epochs with gradient checkpointing. Learning rates and batch sizes varied by model (Table 2), chosen based on empirical constraints of the available GPU memory. Experiments ran on an NVIDIA RTX PRO 6000 Blackwell Server Edition GPU (standard models) and an NVIDIA A100 40GB (Longformer). In addition to individual models, an ensemble by soft-voting of the three best models by macro-F1 (BERTimbau-large, BERTimbau-base, and RoBERTaLexPT-base) with equal weights was evaluated to represent the peak capability of this family. Since about 79.4% of passages exceed 512 tokens, truncation discards part of the text in the standard models and may favor the start of each passage, a potential lead bias examined in Section 4. Longformer reads up to $4,096$ tokens natively and avoids this. An explicit chunking strategy was not employed in this study in order to evaluate the models under a strict single-sequence setting, which remains a limitation for long legal texts. All encoders were trained with a single random seed, leaving run-to-run variance unestimated.

Table 2. Fine-tuning configuration per BERT model.

Model	Batch	Learning rate	Warmup	Max len
BERTimbau-base	128	4.0×10^{-5}	6%	512
BERTimbau-large	48	4.9×10^{-5}	6%	512
RoBERTaLexPT-base	128	4.0×10^{-5}	6%	512
RoBERTaCrawlPT-base	128	4.0×10^{-5}	6%	512
XLM-R-base	96	4.0×10^{-5}	6%	512
mBERT	128	4.0×10^{-5}	6%	512
mDeBERTa-v3-base	96	2.0×10^{-5}	10%	512
Longformer-base	8	3.0×10^{-5}	10%	4096

Large language models. Eight LLMs were evaluated via inference: Claude Sonnet 4.6 [Anthropic 2026], Claude 3 Haiku [Anthropic 2024], Qwen3 32B [Yang et al. 2025], Mistral Large 3 [Mistral AI 2025], Gemma 3 27B [Gemma Team et al. 2025], DeepSeek V3.2 [DeepSeek-AI et al. 2025], and Magistral Small 2509 [Mistral AI et al. 2025], accessed via Amazon Bedrock, as well as Sabiá-4, a Brazilian LLM [Laitz et al. 2026], accessed via the Maritaca API. Each model was run under four prompting strategies: zero-shot (direct instruction), few-shot (one example per class), evidence-constrained (output

plus a supporting excerpt), and decomposed (intermediate questions combined by a priority rule in code). Inference used temperature 0.0, up to $1,024$ output tokens, and one call per instance; every passage fitted within each model's context window without truncation. A parser reads the label from the JSON response, falling back to keyword search. Responses that could not be mapped to a valid class were counted as errors rather than discarded, to reflect an automated workflow, and the parsing failure rate was recorded per model and strategy. The evaluation did not use provider-enforced JSON schemas or tool calls, which simplifies the comparison but may understate the reliability LLMs could reach.

Experimental protocol and metrics. The stratified split used 80% for training, 10% for validation, and 10% for testing, giving $20,377$, $2,546$, and $2,559$ passages. The test set (877 fact, 814 legal thesis, 868 noise) was used by all approaches: classical and BERT models were trained on this split, and the LLMs were run on the same test set by inference. Since the two pipelines were built separately, the LLMs were re-evaluated on the reference test set to keep the comparison identical. Evaluation uses accuracy, macro precision, macro recall, macro-F1, per-class F1, and the confusion matrix, with MCC and Cohen's kappa as complementary metrics. Macro-F1 is the primary metric because it weights the three classes equally; it is the mean of $F1_c = 2P_cR_c/(P_c + R_c)$, where P_c and R_c are the precision and recall of class c. A 95% confidence interval for macro-F1 comes from a percentile bootstrap with $2,000$ resamples and a fixed seed, and pairwise differences are tested with the exact McNemar test on per-instance predictions. Each run recorded its configuration and environment with fixed seeds for reproducibility; classical models used `scikit-learn`, `xgboost`, and `lightgbm`, BERT models used `transformers`, and the LLMs were accessed via Amazon Bedrock and the Maritaca API. Code, configurations, and per-instance predictions are publicly available, including the best fine-tuned model on HuggingFace[1]. The BERT models used the Hugging Face `transformers`[2] API for Python.

4. Results and Discussion

Overall performance. Table 3 presents the best configuration of each family on the test set. The three top configurations of the BERT family are presented to contrast the ensemble, the best single model, and the long-context architecture. BERT-based models led with macro-F1 close to 0.99; the ensemble reached 0.990, with BERTimbau-large alone statistically indistinguishable (McNemar $p = 1.0$). Longformer reached 0.983, and the best classical model (LightGBM) 0.973; the gap between them is significant (McNemar $p < 10^{-6}$). Among LLMs, Qwen3-32B evidence-constrained was best (0.853). The non-overlapping confidence intervals of the BERT ensemble and LightGBM confirm the cross-family difference; Longformer's gap to BERTimbau-large (McNemar $p = 0.002$) shows that extended context did not compensate for the lack of Portuguese-specific pretraining. We note that the gap between the ensemble and BERTimbau-large is only 25 versus 13 errors out of $2,559$, which is small enough that single-seed variance could reorder the top models.

[1]Repository: https://github.com/Willgnner-Santos/fact-claim-noise-classifier. Model: https://huggingface.co/Willgnner-Santos/ftr-bertimbau-large.

[2]https://huggingface.co/docs/transformers/

Impact of anonymization markers. To test whether the high concentration of anonymization placeholders in the noise class gave the models a lexical shortcut, we re-evaluated BERTimbau-large on the test set after removing every placeholder. Macro-F1 did not fall (0.990 to 0.991, with total errors dropping from 25 to 22), and noise F1 stayed at 0.999. The markers therefore do not explain the high overall scores; the model relies on the surrounding text context.

Table 3. Overall performance of the best configurations (test set, $2,559$ passages), with 95% bootstrap CI for macro-F1.

Approach	Acc.	Prec. M	Rec. M	F1 M	95% CI
BERT – ensemble (3 models)	0.991	0.990	0.991	0.990	[0.987, 0.994]
BERT – BERTimbau-large	0.990	0.990	0.990	0.990	[0.986, 0.994]
BERT – Longformer-base-4096	0.983	0.983	0.983	0.983	[0.977, 0.988]
Classical ML – LightGBM	0.973	0.973	0.973	0.973	[0.967, 0.979]
LLM – Qwen3-32B (evidence)	0.850	0.876	0.852	0.853	[0.839, 0.866]

The scores are high and should be read with caution, given the superficial cues, structured annotation criteria, and standardized formats discussed in the methodology. One consequence is already visible: even Longformer, which reads complete documents, did not surpass the fine-tuned 512-token models, which indicates that the discriminating signals lie mostly in the first tokens of each passage.

Per-class performance and error analysis. Table 4 shows F1 by class. For BERT-based and classical models, noise was the easiest class (F1 close to 1.0), reflecting regular patterns and layout cues. Fact and legal thesis were harder. For LLMs, patterns vary; Qwen3-32B struggled with legal thesis (F1 0.812) but handled noise better (0.893). However, several other LLMs found noise to be among the hardest, struggling to recognize administrative content without fine-tuning. Figure 2 shows the BERT ensemble's errors are concentrated between fact and legal thesis (16 fact as legal thesis, 7 reverse). This is sensitive legally since both appear interleaved. Longformer showed the same profile, with near-perfect noise (1 error in 868). A closer look at BERTimbau-large's 25 errors reveals structurally motivated confusions: explicit legal references embedded in factual narration, or legal arguments dominated by concrete opening details. These patterns match the human evaluation (Section 5).

Table 4. Per-class F1-score for the main approaches (test set).

Approach	Fact	Legal thesis	Noise
BERT – ensemble (3 models)	0.987	0.985	0.999
BERT – BERTimbau-large	0.986	0.985	0.999
BERT – Longformer-base-4096	0.977	0.974	0.997
Classical ML – LightGBM	0.962	0.959	0.999
LLM – Qwen3-32B (evidence)	0.853	0.812	0.893

To check whether truncation at 512 tokens hurts longer texts, Table 5 reports BERTimbau-large macro-F1 by length range. Performance is stable across ranges, including passages over $1,024$ words ($n = 1,085$; macro-F1 $= 0.985$). If the model relied only on the start of the text, longer ranges would drop, which does not happen. This weakens the severe-lead-bias hypothesis, though it does not completely rule it out since alternative chunking or end-truncation strategies were not tested; key information in long documents may still sit in opening passages, which is consistent with Longformer's modest gain on long passages.

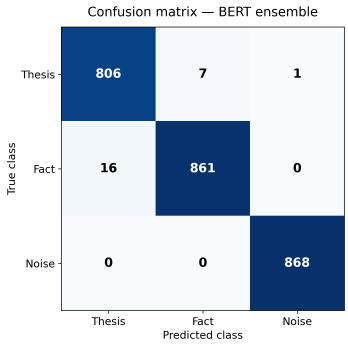


Figure 2. Confusion matrix of the best overall model (BERT ensemble).

Table 5. BERTimbau-large macro-F1 by passage length range (in words), test set.

Range (words)	n	Macro-F1
0–256	415	0.986
256–512	527	0.989
512–1,024	532	0.992
$>$ 1,024	1,085	0.985

Cross-family comparison. Classical models were competitive: LightGBM, linear SVM, and Logistic Regression all exceeded 0.95 macro-F1, which points to strong lexical patterns that TF-IDF captures well. They lose ground when meaning depends on long-range context or word order. XGBoost was the exception (0.642), likely due to its sensitivity to the high dimensionality of sparse representations. The classical models were not tuned with a systematic search, but the gap from LightGBM to BERTimbau-large is significant (McNemar $p < 10^{-6}$, 56 versus 13 errors), which bounds how much a larger search could gain. BERT-based models performed best, in line with their ability to capture bidirectional context and semantic relations, and the three best were all pretrained on Portuguese. Longformer, pretrained on English and fine-tuned on this corpus, reached 0.983, above the classical models but below the Portuguese-pretrained encoders. This gap suggests that language-specific pretraining matters more than extended context here, since the discriminating features are reachable within the first 512 tokens and performance is stable across length ranges (Table 5), so truncation was not the limiting factor. A Portuguese-pretrained long-context encoder would test this more directly and is left as future work.

The LLMs, evaluated without fine-tuning, fell below both other families. This substantial gap likely occurred because out-of-the-box LLMs struggled with the specific legal dialect, formatting, and institutional definitions of Brazilian petitions, lacking the task-specific alignment provided by fine-tuning. The best usable result was Qwen3-32B (evidence-constrained), with macro-F1 of 0.853 and a very low format failure rate (0.1%). Claude 3 Haiku with few-shot reached a nominally higher F1 (0.855), but 13.6% of its responses had an invalid format and were counted as errors, which shows that formatting constraints alone can heavily shift performance. The families also differ in cost. Classical models are cheap to train and run on CPU; BERT models need a GPU for fine-tuning, but training was fast (minutes per 512-token encoder, about 3h for Longformer on an A100) and inference is predictable; LLMs carry a per-document cost and depend on external services. Figure 3 relates LLM inference cost to macro-F1, which ranged from about US$\$\,0.41$ to US$\$\,12.88$ per thousand documents. Model selection should not rely on the top metric alone: in a high-volume workflow, cost, stability, privacy, maintenance,

and auditability also matter. Under these criteria, and within the limits of this study, the Portuguese-pretrained encoders offered the best balance, with BERTimbau-large alone close to the ensemble and simpler to run. Classical models remain a viable low-cost baseline, and one practical direction is a hybrid setup where classical or BERT models handle the bulk of the classification and LLMs are reserved for ambiguous cases.

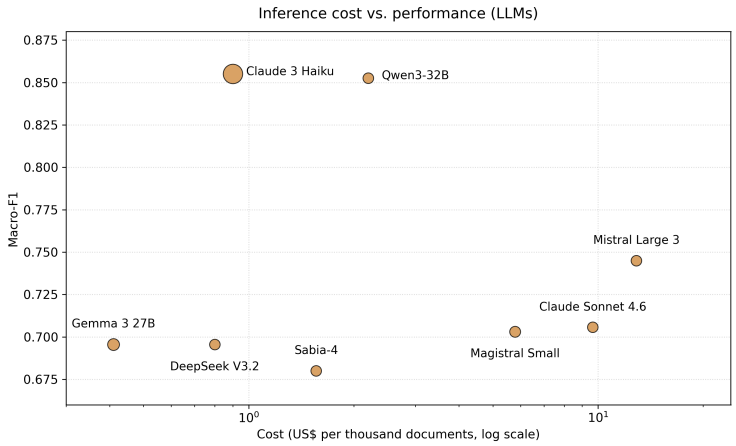


Figure 3. Relationship between inference cost and performance for LLMs. Point size indicates the invalid-format response rate.

5. Human Evaluation

To assess task difficulty, three legal professionals independently classified 10 test passages: six predicted correctly by BERTimbau-large with high confidence (≥ 0.99), two per class, and four model errors (fact/legal-thesis pair). Reviewers chose a class with confidence (1 to 5). No model predictions were shown. The reviewers agreed on six passages and diverged on the four model errors (Fleiss' $\kappa = 0.46$, moderate). Human-model agreement was 70%. Mean confidence was 4.7 on unanimous passages against 3.5 when a reviewer diverged, indicating genuine ambiguity. Noise passages drew full agreement and highest confidence (5.0), matching the automated models. The moderate agreement confirms that hard cases reflect genuine ambiguity rather than a model artifact. While this hand-picked subset cannot establish an upper bound, it motivates future work on corpus-scale annotation and a human-AI triage workflow.

6. Limitations and Threats to Validity

These findings should be interpreted with care. Although anonymization markers correlate with class labels, ablation suggests they do not drive performance. Exact-match deduplication was performed, but near-duplicate analysis was limited to a 2,000-passage training sample without per-class stratification; residual score inflation for fact and legal-thesis classes cannot be fully excluded. Classical baselines were not systematically tuned, though the large gap to BERTimbau-large suggests tuning alone would not close it. All encoders used a single random seed, leaving run-to-run variance unestimated. LLMs were evaluated out-of-the-box without fine-tuning or structured outputs, which may underestimate their reliability. Longformer was pretrained in English; a Portuguese long-context encoder would better isolate context length from pretraining language. Finally, the corpus comes from a single institution, and human evaluation covered only 10 passages and 3 reviewers. Future work should address these through broader agreement analysis, corpus-scale validation, and production-oriented evaluation.

7. Conclusion

This work compared classical models, BERT-based encoders (including Longformer), and LLMs for classifying initial petition passages into fact, legal thesis, and noise using a single Brazilian Portuguese corpus. Portuguese-pretrained encoders achieved the strongest performance (macro-F1 ~ 0.99), followed by Longformer, classical models, and LLMs without fine-tuning. Longformer did not surpass the 512-token models, suggesting language-specific pretraining matters more than extended context here, and that discriminative signals often appear early in each passage. Error analysis identified passages mixing factual narration and legal argument as the main failure mode, confirmed by human evaluation ($\kappa = 0.46$) as genuine ambiguity rather than a model artifact. While these approaches support document triage, findings should be interpreted with care since structured annotation criteria may introduce predictable lexical patterns. Institutional adoption should consider performance, cost, stability, privacy, maintenance, and auditability. Future work includes corpus-scale annotation agreement, testing a Portuguese-pretrained long-context encoder, and real-world triage validation.

References

Adsul, N. and Deshmukh, P. (2023). Automatically classify each sentence from Indian legal judgements to its semantic role. In *Proceedings of the 7th International Conference on Computer Applications in Electrical Engineering (CERA 2023)*. IEEE.

Anthropic (2024). The claude 3 model family: Opus, sonnet, haiku. https://www-cdn.anthropic.com/de8ba9b01c9ab7cbabf5c33b80b7bbc618857627/Model_Card_Claude_3.pdf.

Anthropic (2026). System card: Claude sonnet 4.6. https://anthropic.com/claude-sonnet-4-6-system-card.

Aragy, R., Fernandes, E. R., and Caceres, E. N. (2021). Rhetorical role identification for Portuguese legal documents. In *Intelligent Systems (BRACIS 2021)*, Lecture Notes in Computer Science. Springer.

Belfathi, A., Hernandez, N., and Monceaux, L. (2023). Enhancing pre-trained language models with sentence position embeddings for rhetorical roles recognition in legal opinions. In *CEUR Workshop Proceedings*.

Beltagy, I., Peters, M. E., and Cohan, A. (2020). Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Bendová, K., Knap, T., Černý, J., Pour, V., Savelka, J., Kvapilíková, I., and Drápal, J. (2025). What are the facts? automated extraction of court-established facts from criminal-court opinions. In *CEUR Workshop Proceedings*.

Bhattacharya, P., Paul, S., Ghosh, K., Ghosh, S., and Wyner, A. (2023). DeepRhole: Deep learning for rhetorical role labeling of sentences in legal case documents. *Artificial Intelligence and Law*.

Bomfim, F., Neto, J., Neves, A., Furtado, V., Silva, S., and Pinheiro, V. (2026). Automating legal reasoning with generative AI: A framework for judicial decision support. In *Proceedings of the 20th International Conference on Artificial Intelligence and Law (ICAIL 2025)*.

Bonfim, F. C. J., Silva, S. P., Neves, A. S., Santos, H., Neto, J. A. M., Araujo, J. L. B., Pires, R. S., Costa, R. A. D. S., Furtado, V., and Pinheiro, V. (2026). SARA: Leveraging LLM agents and jurisprudential ontologies for automated legal reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Canaverde, B., Pires, T. P., Ribeiro, L. M., and Martins, A. F. (2025). Legalbench. pt: a benchmark for portuguese law. *arXiv preprint arXiv:2502.16357*.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2019). Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Conselho Nacional de Justiça (Brasil) (2025). Resolução nº 615, de 11 de março de 2025. *Atos Normativos do Conselho Nacional de Justiça*. Access em: 9 dez. 2025.

Csányi, G. M., Üveges, I., Lakatos, D., Ripszám, D., Kozák, K., Nagy, D., and Vadász, J. P. (2025). Sentence-level rhetorical role labeling in judicial decisions. *Big Data and Cognitive Computing*.

Dal Pont, T. R., Billi, M., Sabo, I. C., Lagioia, F., Hübner, J. F., Sartor, G., and Rover, A. J. (2026). Factor extraction from pretrial detention decisions by Italian and Brazilian supreme courts: A knowledge graph perspective. *Computer Law and Security Review*.

Darji, H., Heckelmann, M., Kratsch, C., and De Melo, G. (2025). Segmentation and processing of German court decisions from open legal data. In *Frontiers in Artificial Intelligence and Applications*.

DeepSeek-AI, Liu, A., et al. (2025). Deepseek-v3.2: Pushing the frontier of open large language models. *arXiv preprint arXiv:2512.02556*. Model used via Amazon Bedrock.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

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.

Gemma Team, Kamath, A., et al. (2025). Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*. Reference for Gemma 3 27B. Model used via Amazon Bedrock.

He, P., Gao, J., and Chen, W. (2021a). Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing.

He, P., Liu, X., Gao, J., and Chen, W. (2021b). Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

Kalamkar, P., Tiwari, A., Agarwal, A., Karn, S., Gupta, S., Raghavan, V., and Modi, A. (2022). Corpus for automatic structuring of legal documents. In *Proceedings of the Language Resources and Evaluation Conference (LREC 2022)*.

Lai, J., Gan, W., Wu, J., Qi, Z., and Yu, P. S. (2024). Large language models in law: A survey. *AI Open*, 5:181–196.

Laitz, T., Sales Almeida, T., Abonizio, H., Malaquias Junior, R., Bon'as, G. K., Piau, M., Larcher, C., Pires, R., and Nogueira, R. (2026). Sabi'a-4 technical report. *arXiv preprint arXiv:2603.10213*. Model accessed via Maritaca API.

Liepin, R., Galloni, F., Lagioia, F., Lippi, M., Musicco, M., Sayin, B., Passerini, A., and Sartor, G. (2025). Legal argument mining: Recent trends and open challenges. In *CEUR Workshop Proceedings*.

Lou, A., Salaün, O., Westermann, H., and Kosseim, L. (2021). Extracting facts from case rulings through paragraph segmentation of judicial decisions. In *Legal Knowledge and Information Systems*, Lecture Notes in Computer Science. Springer.

Malik, V., Sanjay, R., Guha, S. K., Hazarika, A., Nigam, S., Bhattacharya, A., and Modi, A. (2022). Semantic segmentation of legal documents via rhetorical roles. In *Proceedings of the Natural Legal Language Processing Workshop (NLLP 2022)*.

Mistral AI (2025). Mistral large 3. https://docs.mistral.ai/models/model-cards/mistral-large-3-25-12.

Mistral AI, Rastogi, A., Jiang, A. Q., Lo, A., Berrada, G., Lample, G., et al. (2025). Magistral. *arXiv preprint arXiv:2506.10910*. Reference for Magistral Small 2509. Model used via Amazon Bedrock.

Mohan, D. and Nair, L. R. (2024). Deep learning-based semantic segmentation for legal texts: Unveiling rhetorical roles in legal case documents. In *E3S Web of Conferences*.

Mok, J. R., Mok, W. Y., and Mok, R. V. (2021). Sentence classification for contract law cases: A natural language processing approach. In *Proceedings of the 18th International Conference on Artificial Intelligence and Law (ICAIL 2021)*.

Östling, A., Sargeant, H., Xie, H., Bull, L., Terenin, A., Jonsson, L., Magnusson, M., and Steffek, F. (2023). The cambridge law corpus: A dataset for legal ai research. *Advances in Neural Information Processing Systems*, 36:41355–41385.

Rapôso, C. F. L., de Lima, H. M., de Oliveira Junior, W. F., Silva, P. A. F., and de Souza Barros, E. E. (2019). Lgpd-lei geral de proteção de dados pessoais em tecnologia da informação: Revisão sistemática. *RACE-Revista de Administração do Cesmac*, 4:58–67.

Santos, W. F., Galvão Filho, A. R., de Oliveira, S. S. T., and Presa, J. P. C. (2025). Comparing prompt-based llms, fine-tuning, and classical models for legal text classification in portuguese. In *Encontro Nacional de Inteligência Artificial e Computacional (ENIAC)*, pages 1138–1149. SBC.

Santos, W. F., Viana, J. G. G., Júnior, A. P. d. C., Trindade, F. R., and Silva, N. F. F. d. (2026). LexIris-pt and LexBert-pt: Specialized sentence embeddings for legal similarity in Brazilian Portuguese. In Souza, M., de Dios-Flores, I., Santos, D., Freitas, L., Souza, J. W. d. C., and Ribeiro, E., editors, *Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026) - Vol. 1*, pages 540–550, Salvador, Brazil. Association for Computational Linguistics.

Šavelka, J. and Ashley, K. D. (2018). Segmenting U.S. court decisions into functional and issue specific parts. In *Frontiers in Artificial Intelligence and Applications (JURIX 2018)*.

Souza, F., Nogueira, R., and Lotufo, R. (2020). BERTimbau: Pretrained BERT models for Brazilian Portuguese. In *Intelligent Systems (BRACIS 2020)*, volume 12319 of *Lecture Notes in Computer Science*, pages 403–417. Springer.

Tomasi, S. (2026). Does ChatGPT have law sense? reflections on an experimental study of argument mining in judicial decisions of the Italian court of cassation. *BioLaw Journal*.

Yang, A. et al. (2025). Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Zaheer, M., Guruganesh, G., Dubey, A., Ainslie, J., Alberti, C., Ontañón, S., Pham, P., Ravula, A., Wang, Q., Yang, L., and Ahmed, A. (2020). Big bird: Transformers for longer sequences. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33.

A. LLM Prompt Templates

The four prompt templates used for LLM inference are presented below. Variables such as `{texto}` and `{exemplo_fato}` were injected programmatically. All prompts requested JSON output.

1. Zero-shot

"Classifique o trecho abaixo em uma das três categorias: 0 - Tese Jurídica, 1 - Fato, 2 - Ruído (cabeçalhos, rodapés, dados processuais). Trecho: `{texto}`. Responda em formato JSON: `{"classe": número}`"

2. Few-shot

"Siga os exemplos para classificar o trecho em: 0 - Tese, 1 - Fato, 2 - Ruído. Exemplo Fato: `{exemplo_fato}`. Exemplo Tese: `{exemplo_tese}`. Exemplo Ruído: `{exemplo_ruido}`. Trecho: `{texto}`. Responda em formato JSON: `{"classe": número}`"

3. Evidence-constrained

"Classifique o trecho em: 0 - Tese, 1 - Fato, 2 - Ruído. Justifique copiando uma evidência exata do texto. Trecho: `{texto}`. Responda em formato JSON: `{"classe": número, "evidencia": "texto copiado"}`"

4. Decomposed

"Analise o trecho: `{texto}`. Responda em formato JSON:

```
{"narra_evento_concreto": true/false,
"cita_lei_ou_jurisprudencia": true/false,
"eh_apenas_cabecalho_ou_dado_processual": true/false}"
```