Representation Engineering in the Legal Domain: A Comparative Study of Linear and Non-Linear Probing for Initial Petition Classification

Pedro Inácio C. de A. Soares[1], **Antonio F. L. Jacob Jr.**[2], **Gustavo S. Silva**[2]

[1]Curso de Engenharia da Computação – Universidade Estadual do Maranhão (UEMA)
São Luís – MA – Brasil

[2]Programa de Pós-Graduação em Engenharia da Computação e Sistemas (PECS)
Universidade Estadual do Maranhão (UEMA)

pedro.20210003059@aluno.uema.br, gustavo.20241004427@aluno.uema.br

Abstract. *This work compares linear and non-linear probing applied to the hidden states of the Jurema-7B model to classify Repetitive Demand Resolution Incident (IRDR) petitions. To mitigate textual length and class imbalance, summarization strategies and context injection via hybrid RAG were integrated. Results indicate that non-linear probing outperforms the linear approach, achieving a macro F1-score of 0.617, with statistically significant differences confirmed by McNemar's test with Holm-Bonferroni correction in seven of the eight evaluated scenarios. We conclude that non-linear representation engineering is a viable alternative for automating procedural screening in the Brazilian judiciary.*

1. Introduction

The Brazilian Judiciary faces a structural productivity crisis. In April 2026, pending lawsuits exceeded 76 million, with a 63% congestion rate [Conselho Nacional de Justiça (CNJ) 2026]. A considerable portion of this volume consists of initial petitions, documents through which plaintiffs file lawsuits and present facts and legal grounds [Chagas et al. 2025], and their constant influx demands ongoing screening by court staff. To manage this workflow, the 2015 Civil Procedure Code created the Repetitive Demand Resolution Incidents (IRDR), which group lawsuits involving the same legal issue for joint judgment, seeking to standardize case law [Tupinambá and Teixeira 2022]. Automatically classifying petitions into IRDR categories can thus reduce screening time and enhance consistency in decisions.

Automating this type of judicial task faces inherent challenges for natural language processing (NLP), including long texts with thousands of tokens and specialized terminology that limit the generalization of models trained on generic domains [da Silva Junior et al. 2025]. The problem is intensified by the natural imbalance of legal cases. In IRDRs, for example, majority classes may contain nearly all records, since mass litigation topics, such as banking contracts and social security benefits, produce far more cases than less frequent issues. To address these limitations, large language models (LLMs) are a promising solution when combined with representation engineering, which analyzes internal model representations without changing network weights [Bartoszcze et al. 2025].

Within this scope, the probing technique stands out, training classifiers directly on the LLM's activation vectors (hidden states). Its effectiveness depends on the linear separability of data in the representation space, where linear classifiers are computationally simpler, while non-linear ones can decode more complex abstractions [Jin and Rinard 2024]. In Portuguese legal NLP, benchmarks and architectural evaluations remain scarce [Carvalho Neto 2026], so the choice between linear and non-linear probing is still unresolved. In this domain, structural complexity and highly specialized vocabulary [do Carmo 2024] likely shape how knowledge is organized in the LLM's internal layers [Sawtell et al. 2024].

This work compares linear probing and non-linear probing applied to the hidden states of a specialized legal LLM to classify initial petitions from Court of Justice of the State of Maranhão (TJMA) into IRDR categories. The study also analyzes the impact of summarization techniques on the quality of extracted representations and the effect of exemplar retrieval on these vectors, with statistical validation of the results. In practice, the proposed pipeline applies representation engineering as a viable alternative to support procedural screening in the Judiciary, with the potential to reduce manual classification time and contribute to decision-making equality across Brazilian courts.

2. Related Work

Since language models based on the Transformer architecture were introduced, large language models (LLMs) have been widely used in the NLP field. However, for more complex domains, such as the legal field, traditional prompt engineering methods, such as zero-shot, few-shot, and Chain-of-Thought (CoT), show inconsistent performance in fine-grained legal classification tasks when restricted to these purely generative approaches [Hakimi Parizi et al. 2023, Jayakumar et al. 2023]. To overcome the vulnerabilities of textual generation and access the latent knowledge of these architectures, the literature has turned to representation engineering. Within this context lies the probing technique, which relies on freezing network weights to extract knowledge directly from activation matrices [Bartoszcze et al. 2025], using classifiers called probes [Belinkov 2022].

Research shows that information from the hidden states of language models possesses an abstraction power that helps to understand facts more effectively than the final response generated by the models [Azaria and Mitchell 2023, Mallen et al. 2024]. Utilizing the internal state of LLMs enables an understanding of how the Transformer architecture processes complex data [Garcia et al. 2025], such as the technical vocabulary and argumentative logic of the legal domain. In this sense, the study by [Skean et al. 2025] found that intermediate layers outperform the final layer in accuracy by up to 16%, as the latter is optimized for next-word generation rather than abstraction, a conclusion also reached by [Aldabergen et al. 2026], who argue that final layers are less suited to capturing complex abstractions.

The choice of classifier trained on model activations is important for the probing task. Under the Linear Representation Hypothesis, where high-level concepts correspond to linear directions in representation space [Park et al. 2024], linear probes (single-layer classifiers) work well and can outperform complex probes when features are already linearly separable, as in academic essays [Chi et al. 2025]. However, when decoding complex, high-dimensional logical information, non-linear probes (neural networks with hid-

den layers) are preferable, since such information may not be linearly separable in language model activations [Li et al. 2024]. In the English legal domain, non-linear architectures trained on activations outperform generative methods [Hakimi Parizi et al. 2023]. Accurate activation extraction also requires precise model context, which can be optimized by methods such as hybrid Retrieval-Augmented Generation (RAG) for selecting examples [Akarsu et al. 2026].

In the Brazilian context, probing is employed as a means of semantic auditing for Portuguese models [Peixoto et al. 2026, Wilkens et al. 2023]. However, national scientific production lacks investigations into how Brazilian legal data behaves regarding linearity and complexity. Long documents, such as initial petitions, create obstacles to the information processing of LLMs. This bottleneck makes the use of summarization strategies essential, incorporating language models specialized in the legal domain, such as BumbaBERT [1] [do Carmo 2024], to reduce text length [Araújo et al. 2026].

The present research aims to fill the gap in studies on the linear separability of Portuguese legal representations and to evaluate the impact of summarization strategies on long documents. To this end, the study contrasts linear and non-linear probing in the intermediate (17th) and final (28th) layers of the quantized Jurema-7B[2] model. By combining summarization techniques and hybrid RAG, it aims to determine whether legal texts from TJMA can be categorized effectively through linear methods or if more complex, non-linear approaches are required.

3. Methodology

To ensure that the work aligns with the real needs of the judiciary and with machine learning experimentation, the project's lifecycle was structured based on the Cross-Industry Standard Process for Data Mining (CRISP-DM) reference model [Wirth and Hipp 2000]. The approach was divided into the following phases: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment.

3.1. Business and Data Understanding

This research focuses on initial petitions extracted from a Brazilian court. To delimit the problem, the dataset was filtered to contain only petitions classified as Repetitive Demand Resolution Incidents (IRDR). The objective is to classify these documents based on the seven active categories in the court: IRDR-1, IRDR-2, IRDR-3, IRDR-4, IRDR-5, IRDR-7, and IRDR-8. The main challenge regarding this dataset lies in the severe class imbalance and the substantial length of the documents.

Table 1 shows a strong imbalance, as the IRDR-1 class concentrates 65.80% of the records, while IRDR-2 and IRDR-7 together represent only 0.53%. Beyond this quantitative disparity, the dataset also poses linguistic separability issues. Classes IRDR-1 and IRDR-2, for instance, differ only in whether the salary readjustment is complementary, which can create textual ambiguity. This semantic and structural proximity reduces linear separability, favors pattern confusion by classifiers, and helps explain the difficulty in distinguishing classes.

[1] https://huggingface.co/fabricioalmeida/bumbabert-small-v3
[2] https://huggingface.co/Jurema-br/Jurema-7B

Table 1. Distribution and anonymized description of IRDR classes

Class	Theme / Legal Description	Records	Percentage
IRDR-1	State civil servants' salary readjustment	3,571	65.80%
IRDR-5	Payroll deduction loans	1,058	19.49%
IRDR-3	Surplus candidates in public service examinations	470	8.66%
IRDR-8	Statute of limitations on military promotions	248	4.57%
IRDR-4	Bank fees for INSS beneficiaries	51	0.94%
IRDR-2	Complementary salary readjustment for state civil servants	23	0.42%
IRDR-7	Legal fees in individual enforcement proceedings	6	0.11%
Total		**5,427**	**100.00%**

3.2. Data Preparation

To ensure the petitions fit within the context window limit of the models used in the experiments, the dataset was processed with two parallel summarization methods, each constrained to 512 tokens, in order to compare the effects of maintaining the original textual structure versus focusing on semantic content extraction.

The first summarization was performed using Head+Tail truncation, which consists of preserving the initial and final portions of the text [Chen and Lv 2022], based on the hypothesis that the qualification of the parties and the procedural conclusion are located in these sections of the document [Almeida et al. 2026]. In contrast, the second approach employed a semantic bias using the BumbaBERT extractive model, pre-trained on Brazilian judiciary data. This algorithm extracts sentences with the highest semantic relevance, creating a summary through extractive summarization, in which no words or facts are altered.

3.3. Modeling

To obtain a semantically rich context for extracting the model's internal state, a hybrid RAG system was implemented. For each summarized petition selected as a target, lexical matching was applied using the BM25 algorithm, widely used for retrieving legal terminologies and Brazilian court documents [da Silva Junior et al. 2025]. This technique was combined with dense searches using the BGE-M3 model and FAISS indexing, merging both rankings via Reciprocal Rank Fusion (RRF).

To prevent label-leakage during cross-validation, the retrieval pool was restricted to the training fold of each target petition. Finally, near-duplicates were filtered using a cosine similarity threshold of 0.95, yielding the three most relevant and distinct examples to compose the dynamic three-shot prompt. The template used the text presented in Listing 1 as an instruction (originally in Portuguese).

```
You are an expert in jurimetrics. Analyze the initial petition and
classify it strictly into one of these categories: IRDR-1, IRDR-2,
IRDR-3, IRDR-4, IRDR-5, IRDR-7, IRDR-8. Reply ONLY with the class
label.

Classify the following petition: {{summary_of_example_petition_1}}
{{class_1}}

<... two additional retrieved examples, same format ...>
```

```
Classify the following petition: {{target_petition_text}}
```

Listing 1. Prompt template for classifying petitions into predefined IRDR categories.

The probing technique was based on extracting the activation vectors generated at the last token, whose representation aggregates the preceding context [Zhao et al. 2026]. To capture different levels of semantic abstraction, the vectors were extracted from the 17th layer (intermediate), as it retains the semantic essence and filters out redundancies, and from the final 28th layer of Qwen2.5-based architecture, which specializes in predicting the next token [Skean et al. 2025]. The model was quantized to 4 bits due to hardware limitations.

The vectors extracted from Jurema-7B were used as input variables for the classifiers, in which two different architectures were compared. Linear probing was implemented using multinomial logistic regression with a single linear layer, optimized with AdamW and cross-entropy loss. The second classifier, non-linear probing, used a multilayer perceptron (MLP) with two hidden layers, batch normalization, ReLU activation, and Dropout regularization, except for the final layer to prevent random suppression of predictions and to provide the raw logits to the cross-entropy function.

Hyperparameter optimization was conducted with the Tree-structured Parzen Estimator (TPE) via the Optuna framework, with 2-fold cross-validation per trial to ensure computational feasibility (50 trials for linear probing, 100 for non-linear probing). For linear probing, the search focused on the learning rate and the intensity of the L2 penalty. For non-linear probing, the search space comprised the number of neurons per hidden layer (128, 256, or 512) and a Dropout rate between 0.0 and 0.5. The hyperparameters used are available in the GitHub repository[3].

Imbalance in the training partitions was mitigated using SMOTE [Andrade et al. 2025], requiring a minimum of 15 samples per class to generate synthetic neighbors, applied exclusively to the training partitions within the cross-validation loops to prevent data leakage.

3.4. Evaluation and Deployment

The pipeline generated 8 representation engineering scenarios, combining two summarizations, two layers, and two probes. All were evaluated using stratified 5-fold cross-validation to preserve the original class distribution. The primary ranking metric was the macro F1-score, as it penalizes errors in minority classes equally to those in majority classes. As IRDR-7 was excluded from all cross-validation folds due to insufficient samples (Section 5), macro F1 and all other metrics reported in Tables 2 and 3 are computed over the remaining six classes.

Statistical validation of the differences between classifiers was performed on out-of-fold predictions pooled across the five CV folds into a single set per petition. For each scenario, McNemar's test was applied to the out-of-fold predictions of linear and non-linear probing, using the corresponding 2×2 contingency table to evaluate the null hypothesis H_0 that both classifiers exhibit identical error rates. To control the family-wise

[3]https://github.com/PedroSoaress/jurema7b-irdr-probing

error rate across the eight scenarios, the Holm-Bonferroni correction was subsequently applied.

4. Results and Discussion

The performance obtained across all scenarios mapped in the experiments for predicting IRDR classes is shown in Table 2, which presents the metrics Accuracy (Acc), Macro Precision (ma-P), Macro Recall (ma-R), and Macro F1-Score (ma-F1).

Table 2. Performance of classification methods

Method	Summarization	Layer	RAG	Acc	ma-P	ma-R	ma-F1
Linear Probing	BumbaBERT	17th	✗	0.68	0.53	0.54	0.520 ± 0.032
			✓	0.70	0.49	0.57	0.522 ± 0.021
		28th	✗	0.68	0.47	0.51	0.485 ± 0.020
			✓	0.70	0.49	**0.60**	0.538 ± 0.037
	Head+Tail	17th	✗	0.76	**0.56**	0.58	$\mathbf{0.572 \pm 0.038}$
			✓	0.76	**0.56**	0.59	$\mathbf{0.572 \pm 0.025}$
		28th	✗	0.75	0.54	0.59	0.571 ± 0.023
			✓	**0.77**	0.53	0.59	0.556 ± 0.017
Non-Linear Probing	BumbaBERT	17th	✗	0.73	0.58	0.55	0.558 ± 0.032
			✓	0.74	0.61	0.56	0.565 ± 0.043
		28th	✗	0.71	0.58	0.54	0.545 ± 0.021
			✓	0.74	0.59	0.57	0.568 ± 0.038
	Head+Tail	17th	✗	**0.79**	0.59	0.59	0.588 ± 0.033
			✓	0.78	0.61	0.59	0.595 ± 0.033
		28th	✗	**0.79**	**0.62**	0.61	$\mathbf{0.617 \pm 0.021}$
			✓	0.78	0.57	**0.63**	0.600 ± 0.010

As seen in Table 2, pipeline configuration and context injection prove as determinant to the outcome as classifier architecture. Non-linear probing outperforms linear probing in every pairing, but the gap between the best and worst scenarios overall is wider than the gap attributable to architecture alone.

Furthermore, the effect of RAG on ma-F1 improves 5 of the 8 configurations, leaves one unchanged (Head+Tail, intermediate layer, linear probing), and reduces ma-F1 in the remaining two, both final-layer Head+Tail configurations, with a 1.7-point drop for non-linear probing (0.617 to 0.600). The largest gains from retrieval occur at the final layer under BumbaBERT summarization with +5.3 points for linear probing (0.485 to 0.538) and +2.3 points for non-linear probing (0.545 to 0.568). Retrieved examples appear to introduce semantic noise mainly when layered onto context that is already well structured positionally.

4.1. Impact of Summarization

The positional Head+Tail summarization outperformed BumbaBERT in all experiments, improving the LLM's hidden state representations across both probing architectures. This likely reflects the formal structure of initial petitions, where party qualifications and procedural requests are concentrated in the opening and closing sections.

By selecting sentences based on semantic weight, BumbaBERT can fragment context. It extracts dense but widely separated excerpts, creating discontinuities in the token

sequences fed to Jurema-7B and disrupting the positional coherence needed for stable hidden-state representations. This reflects a known limitation of similarity-based methods for long documents, where repetitive content is often overemphasized at the expense of rarer yet highly informative excerpts [Araújo et al. 2026, do Carmo 2024].

4.2. Linear and Non-Linear Probing

Non-linear probing outperformed the multinomial logistic regression linear probe in seven of eight configurations; the only exception was the Head+Tail, final-layer, with-RAG setup, where their performance was statistically indistinguishable. Geometrically, this suggests that Jurema-7B's knowledge about the court's procedural incidents is mostly, but not fully, linearly accessible: linear probing captures most of the discriminative signal, while the MLP's residual gains reveal a smaller, statistically detectable non-linear component that its transformations decode more effectively.

The Optuna framework allowed determining the ideal L2 regularization for linear probing and the optimal combination of neurons and Dropout for the non-linear model. SMOTE mitigated the imbalance in the training partitions without inducing overfitting, preserving the original distribution in the validation partitions. The stability and performance impacts in the configuration with Head+Tail summarization are illustrated in Figure 1, which highlights the most significant advantage of the non-linear architecture in the final layer without RAG, with a difference of 4.6 percentage points in ma-F1.

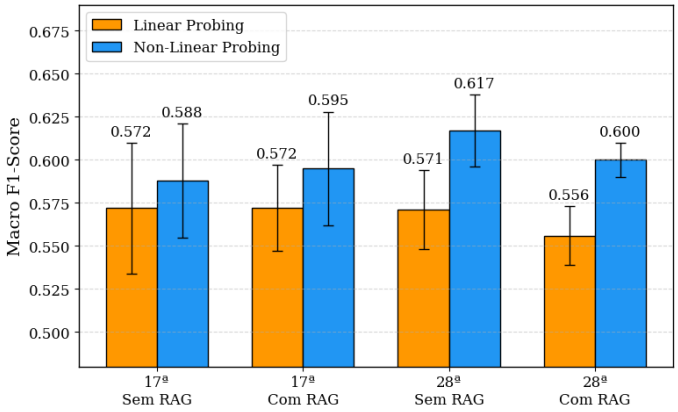


Figure 1. Comparison of methods with positional summarization

It can be observed in the figure that without RAG, final-layer activations expose this non-linearity more clearly, showing the linear classifier's limitation. Figure 2 presents the 2D Principal Component Analysis (PCA) projection of the activation vectors for the configurations with positional summarization.

For the with-RAG scenarios, both the full projection and a zoomed view of the main distribution are shown, since a small and hyper-dense cluster separates from the rest of the data and compresses the visualization of the main distribution when displayed on the same scale.

Manual inspection showed that these outliers are mass-litigation petitions with near-identical boilerplate from one law office, differing only in claimant name and record number. Under RAG, the retrieval repeatedly selects these near-duplicate petitions as exemplars, and the resulting redundancy in the prompt pushes them into an isolated region

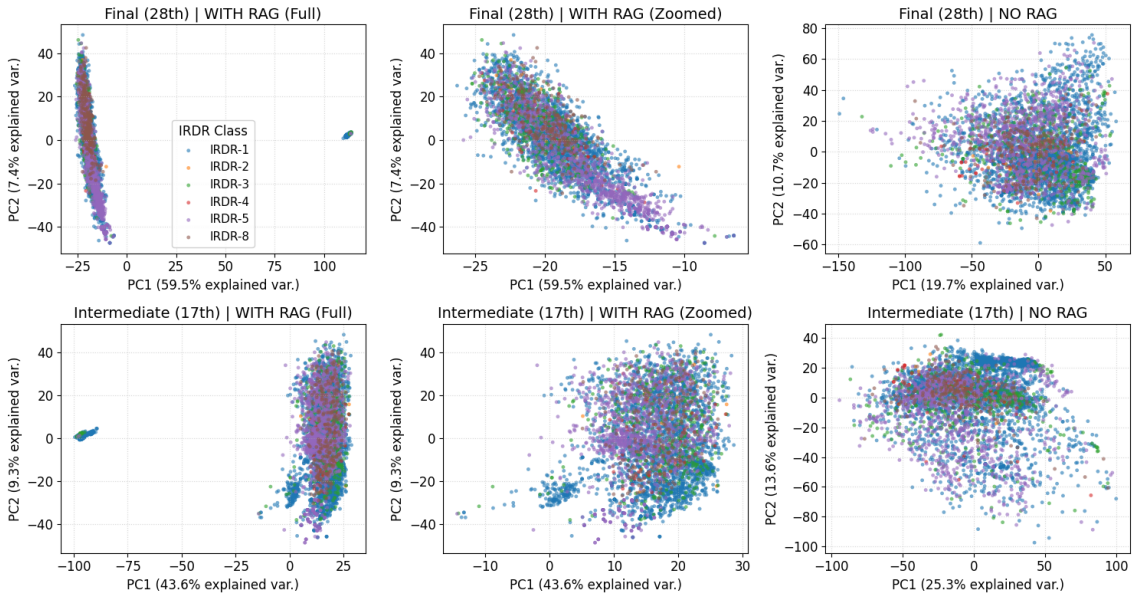


Figure 2. PCA projection of Jurema-7B activation vectors

of latent space. In the zoomed panels, as in the no-RAG scenarios, the IRDR class representations overlap and disperse, qualitatively matching the linear probe's reduced performance. However, formal evidence of non-linear separability comes from the comparative performance and statistical tests in Section 4.3.

Regarding the depth of the activations extracted from the LLM, the final layer outperformed the intermediate layer for non-linear probing in most configurations (3 of 4), with the largest gap observed in the Head+Tail, no-RAG scenario. For linear probing, the opposite tendency emerged, with the intermediate layer generally performing on par with or better than the final layer. In the absence of context retrieval, the non-linear probe applied to final-layer activations with Head+Tail summarization recorded the highest ma-F1 (0.617) among all evaluated configurations. The distribution of predictions for this configuration is shown in Figure 3.

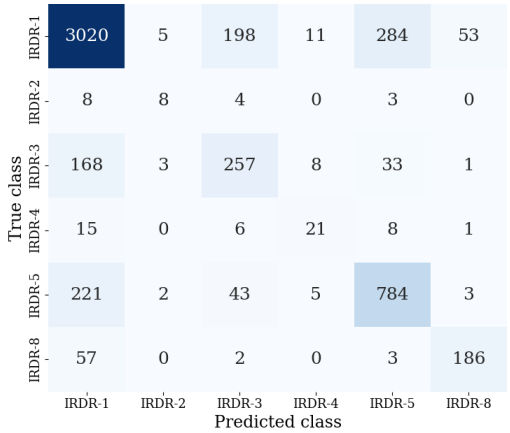


Figure 3. Confusion matrix of the best evaluated model

The confusion matrix shows most correct predictions on the main diagonal, with

residual errors mainly between the majority class (IRDR-1) and IRDR-3/IRDR-5, reflecting their procedural similarity to mass litigation demands. The IRDR-7 class is absent because its small size (6 samples) prevented SMOTE from generating synthetic examples for imbalance handling. Since over 30% of predictions remain misclassified, real-world court deployment will likely require human-in-the-loop review of low-confidence predictions in procedurally similar classes, rather than relying solely on further architectural tuning.

4.3. Statistical Tests

The statistical validation of the differences between linear and non-linear probing is presented in Table 3. Note that the ma-F1 values reported in Table 3 are computed on out-of-fold predictions pooled across all cross-validation folds, as required by the statistical test, and therefore differ slightly from the per-fold averages reported in Table 2.

Table 3. Statistical comparison of linear and non-linear probing

Summarization	**Layer**	**RAG**	**ma-F1**		**McNemar (Holm)**	
			Linear	Non-Linear	ΔF1	p-Holm
BumbaBERT	28th	✓	0.527	0.564	+0.037	<0.00001*
		✗	0.482	0.543	+0.061	<0.00001*
	17th	✓	0.517	0.565	+0.048	<0.00001*
		✗	0.519	0.555	+0.036	<0.00001*
Head+Tail	28th	✓	0.552	0.599	+0.047	0.21130
		✗	0.566	0.614	+0.048	<0.00001*
	17th	✓	0.568	0.595	+0.027	<0.00001*
		✗	0.568	0.587	+0.019	<0.00001*

* Indicates statistically significant difference ($p < 0.05$).

According to the table, non-linear probing outperforms linear probing in seven of eight scenarios ($p < 0.00001$ after Holm-Bonferroni correction), with ma-F1 gains of 1.9–6.1 points. The exception is the Head+Tail, final-layer, with-RAG setup, where LP (0.552) and NLP (0.599) do not differ significantly ($p = 0.211$). This suggests that hybrid RAG structures the final-layer representations well enough for a linear classifier to match the MLP in this setting, making linear probing a more efficient option when both context retrieval and well-structured final-layer representations are available.

5. Limitations

This work presents limitations that should be considered when interpreting the results. The use of SMOTE to mitigate the dataset imbalance may generate less diverse instances for rarer classes, due to the minimum number of k-neighbors. The IRDR-7 class (6 records) was excluded because it lacked the minimum of 15 samples required for 5-fold validation. The Jurema-7B model was quantized to 4 bits for the extraction of activation vectors due to hardware limitations, resulting in a loss of precision compared to the full model. The hyperparameter search via Optuna TPE was performed with 2 folds to make the execution of 100 trials computationally feasible. Despite deduplication filtering, the retrieval mechanism still occasionally selected near-duplicate mass-litigation petitions as exemplars, introducing prompt redundancy.

6. Conclusion

This study evaluated the linear separability of Jurema-7B representations for classifying Brazilian legal documents. Non-linear probing significantly outperformed the linear approach in seven of the eight evaluated configurations after Holm–Bonferroni correction; the only exception was the Head+Tail, final-layer, with-RAG configuration ($p = 0.211$). The best performance was obtained with non-linear probing applied to final-layer representations, Head+Tail summarization, and no RAG, reaching an ma-F1 of 0.617. Together, the classification results and PCA projections indicate that the model encodes a residual, statistically detectable non-linear component that linear probing does not fully capture.

Among summarization strategies, Head+Tail summarization consistently outperformed BumbaBERT, supporting the importance of preserving the positional structure of initial petitions, whose qualifications, legal grounds, and requests tend to occur near the document's beginning and end. Layer depth had a less consistent effect since final-layer representations favored Head+Tail, whereas intermediate-layer representations performed comparably or better with BumbaBERT. These findings support representation engineering as a viable alternative to purely generative classification approaches, particularly when combined with positional summarization and a non-linear probe.

Future work should evaluate petitions from other Brazilian courts, address the scarcity of rare classes such as IRDR-7, apply the pipeline to other legal document types, investigate larger unquantized models, and develop summarization methods that jointly preserve positional structure and semantic relevance.

Acknowledgments and Use of Generative AI

This work was supported by the National Council for Scientific and Technological Development (CNPq), through grant DT-315052/2026-0; by the Court of Justice of Maranhão, through Technical Cooperation Agreement nº 02/2021 (process nº 38328/2020-TJ/MA); and by the Funding Authority for Studies and Projects (FINEP), through the Pró-Amazônia program, project 2.373/24 – CTCCA-II.

It is declared that the generative AI models Claude and NotebookLM were used to assist in revising the manuscript text and refining the research pipeline. Grammarly was used only to review the translation. Authorship and full responsibility for the final content, including plagiarism checks and corrections, rest with the first author.

References

Akarsu, M., Karaman, R. K., and Mierbach, C. (2026). From BM25 to corrective RAG: Benchmarking retrieval strategies for text-and-table documents. arXiv preprint arXiv:2604.01733.

Aldabergen, A., Kynabay, B., and Kadyrov, S. (2026). Layer-wise probing of paralinguistic attributes in fine-tuned Whisper for Kazakh speech. *Engineering, Technology & Applied Science Research*, 16(2):33399–33404.

Almeida, V., S. Pires, R., A. Monteiro Neto, J., and Furtado, V. (2026). Using interpretability to uncover legal petition structures and optimize text classification. In *Proceedings of the Twentieth International Conference on Artificial Intelligence and Law*,

ICAIL '25, page 150–158, New York, NY, USA. Association for Computing Machinery.

Andrade, C. M., Magalhães, G., Asevedo, R., Paes, J., Oliveira, I., Pagano, A., Reis, Z., and Gonçalves, M. A. (2025). Estudo do impacto de dados sintéticos e paráfrases na mitigação do desbalanceamento em tarefas de classificação de textos em português com baixa amostragem. In *Anais do XL Simpósio Brasileiro de Bancos de Dados*, pages 837–843, Porto Alegre, RS, Brasil. SBC.

Araújo, G., Santana, E., and Lobato, F. (2026). Hybrid summarization for Brazilian judicial decisions. In *Anais do XXII Simpósio Brasileiro de Sistemas de Informação*, pages 910–929, Porto Alegre, RS, Brasil. SBC.

Azaria, A. and Mitchell, T. (2023). The internal state of an LLM knows when it's lying. In Bouamor, H., Pino, J., and Bali, K., editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 967–976, Singapore. Association for Computational Linguistics.

Bartoszcze, L., Munshi, S., Sukidi, B., Yen, J., Yang, Z., Williams-King, D., Le, L., Asuzu, K., and Maple, C. (2025). Representation engineering for large-language models: Survey and research challenges. arXiv preprint arXiv:2502.17601.

Belinkov, Y. (2022). Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219.

Carvalho Neto, P. B. d. (2026). LegalBench-BR: A benchmark for evaluating large language models on Brazilian legal decision classification.

Chagas, M. A. d. N., Oliveira, A. F. d., Pinto Netto, M., Pontes, M. d. O. E. d., Alves, C. V. P., and Tizzo, L. G. L. (2025). A estrutura da petição inicial na escrita jurídica: uma abordagem propositiva. *Revista de Ciências Jurídicas e Empresariais*, 26(1):26–37.

Chen, J. and Lv, S. (2022). Long text truncation algorithm based on label embedding in text classification. *Applied Sciences*, 12(19).

Chi, J., Wang, K., Chen, Y., Lin, X., and Xu, Q. (2025). Activations as features: Probing LLMs for generalizable essay scoring representations. arXiv preprint arXiv:2512.19456.

Conselho Nacional de Justiça (CNJ) (2026). Painel de estatísticas do poder judiciário. Acesso em: 17 jun. 2026.

da Silva Junior, D., de Oliveira, D., and Paes, A. (2025). Evaluating text representations for unsupervised legal semantic textual similarity in Brazilian portuguese. *Discover Data*, 3.

do Carmo, F. A. (2024). Representações *Embeddings* Orientadas à Linguagem Jurídica Brasileira. Mestrado em engenharia da computação e sistemas, Universidade Estadual do Maranhão, São Luís/MA.

Garcia, M. H., Couturier, C., Diaz, D. M., Mallick, A., Kyrillidis, A., Sim, R., Ruhle, V., and Rajmohan, S. (2025). Exploring how LLMs capture and represent domain-specific knowledge. arXiv preprint arXiv:2504.16871.

Hakimi Parizi, A., Liu, Y., Nokku, P., Gholamian, S., and Emerson, D. (2023). A comparative study of prompting strategies for legal text classification. In Preoţiuc-Pietro,

D., Goanta, C., Chalkidis, I., Barrett, L., Spanakis, G., and Aletras, N., editors, *Proceedings of the Natural Legal Language Processing Workshop 2023*, pages 258–265, Singapore. Association for Computational Linguistics.

Jayakumar, T., Farooqui, F., and Farooqui, L. (2023). Large language models are legal but they are not: Making the case for a powerful LegalLLM. In Preoțiuc-Pietro, D., Goanta, C., Chalkidis, I., Barrett, L., Spanakis, G., and Aletras, N., editors, *Proceedings of the Natural Legal Language Processing Workshop 2023*, pages 223–229, Singapore. Association for Computational Linguistics.

Jin, C. and Rinard, M. (2024). Latent causal probing: A formal perspective on probing with causal models of data. arXiv preprint arXiv:2407.13765.

Li, K., Hopkins, A. K., Bau, D., Viégas, F., Pfister, H., and Wattenberg, M. (2024). Emergent world representations: Exploring a sequence model trained on a synthetic task. arXiv preprint arXiv:2210.13382.

Mallen, A., Brumley, M., Kharchenko, J., and Belrose, N. (2024). Eliciting latent knowledge from quirky language models. arXiv preprint arXiv:2312.01037.

Park, K., Choe, Y. J., and Veitch, V. (2024). The linear representation hypothesis and the geometry of large language models. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Peixoto, M., Silva, G., Figueredo, G., Silva, P., and Luz, E. J. S. (2026). Global vs. local sentence embeddings for Brazilian Portuguese: Revisiting monolingual models in the age of foundation models. 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 529–539, Salvador, Brazil. Association for Computational Linguistics.

Sawtell, M., Masterman, T., Besen, S., and Brown, J. (2024). Lightweight safety classification using pruned language models. arXiv preprint arXiv:2412.13435.

Skean, O., Arefin, M. R., Zhao, D., Patel, N., Naghiyev, J., LeCun, Y., and Shwartz-Ziv, R. (2025). Layer by layer: Uncovering hidden representations in language models. arXiv preprint arXiv:2502.02013.

Tupinambá, C. and Teixeira, S. T. (2022). O incidente de resolução de demandas repetitivas na justiça do trabalho. *Revista do Tribunal do Trabalho da 2ª Região*, 14(28):50–75.

Wilkens, R., Zilio, L., and Villavicencio, A. (2023). Assessing linguistic generalisation in language models: A dataset for Brazilian portuguese. arXiv preprint arXiv:2305.14070.

Wirth, R. and Hipp, J. (2000). CRISP-DM: Towards a standard process model for data mining. *Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining*.

Zhao, H., He, Z., Tang, Y., Yang, F., Payani, A., Liu, D., and Du, M. (2026). Rep2Text: Decoding full text from a single LLM token representation. arXiv preprint arXiv:2511.06571.