Domain-adaptive T5 for structured information extraction in Brazilian legislative texts using Semantic Role Labels

Frederico Thiers Dutra de Oliveira da Silva[1], **Ana Cristina Bicharra Garcia**[1]

[1]Programa de Pós-Graduação em Informática
Universidade Federal do Estado do Rio de Janeiro (UNIRIO)
Rio de Janeiro – RJ – Brasil

frederico.thiers@edu.unirio.br,cristina.bicharra@uniriotec.br

Abstract. *This paper presents a preliminary investigation into the effectiveness of generative sequence-to-sequence architectures for extracting structured normative information from Brazilian legislative amendments. We evaluate the capacity of a compact, domain-adapted T5 model to map complex legal provisions into a functional Semantic Role Labeling (SRL) schema. By fine-tuning a Portuguese T5-base model on a specialized corpus, our initial results suggest that this text-to-structure approach can reconstruct regulatory intent with higher juridical fidelity than larger, zero-shot general-purpose LLMs. The findings suggests that, in highly conventionalized legal settings, domain-aligned supervision may be a more significant driver for successful extraction than model scale alone. This study provides early evidence for computationally efficient alternatives that preserve institutional drafting patterns, laying the groundwork for more robust scaling in legislative text processing.*

1. Introduction

The automated extraction of structured information from legislative texts is a pivotal task for modern legal engineering [Braz et al. 2018]. As the volume of digital statutory data grows, the ability to convert unstructured legal provisions into machine-readable formats becomes essential for compliance monitoring, policy analysis, and advanced retrieval systems. However, Brazilian legislative language presents a unique challenge: it is characterized by extreme lexical rigidity, complex nested subordinations, and a highly conventionalized drafting style that often eludes general-purpose Natural Language Processing (NLP) models [Vitório et al. 2025].

While traditional legal NLP has primarily relied on sequence labeling for Named Entity Recognition (NER), these methods are often insufficient for capturing the functional logic of a legal rule [Araujo and Silveira 2025]. Identifying a person or an organization is fundamentally different from understanding the prescriptive force of a norm—i.e., who is affected, what is the objective, and under what conditions the rule applies. This study addresses this gap by evaluating the validity of a generative encoder-decoder architecture for mapping Brazilian legislative amendments into structured normative roles.

The primary objective of this research is to verify whether a specialized T5 model, fine-tuned on juridical language, can accurately reconstruct the essential elements of a legal provision. We frame this problem not as a simple classification task, but as a generative text-to-structure mapping by adopting a reduced Semantic Role Labeling (SRL)

schema [Humphreys et al. 2020]. Futhermore we aim to isolate the functional segments of regulatory intent without the heavy structural overhead associated with document-level frameworks like LegalRuleML ([Palmirani et al. 2011]).

This work is of particular interest to the legal tech community as it explores the efficiency of domain-specific models. Our experiments contrast the performance of a compact, fine-tuned PTT5-base model against significantly larger general-purpose LLMs, testing the hypothesis that domain-aligned supervision is the primary driver for successful normative extraction.

The findings presented herein suggests that specialized adaptation effectively compensates for parameter scale in structurally regular legal settings.

2. Related Work

Research in Legal NLP has shifted from surface tasks, such as Named Entity Recognition (NER), toward structural semantic extraction [Araujo and Silveira 2025]. In the Brazilian context, early efforts employed BiLSTM-CRF for procedural NER [Batista et al. 2021], while studies on legislative corpora indicate that transformer-based encoders face limitations in statutory text due to lexical rigidity [Vitória et al. 2025]. Recent surveys confirm that while Large Language Models (LLMs) are effective, they still struggle with complex tasks involving intricate schemas, with well-tuned, domain-specific models often yielding superior results [Deng et al. 2022].

A specialized niche focuses on mapping norms into formal schemas. While frameworks like LegalRuleML ([Athan et al. 2013, Palmirani et al. 2011]) and Akoma Ntoso ([Avgerinos Loutsaris et al. 2023]) provide standards for document-level modeling, their heavy structural overhead is ill-suited for the granular identification of roles within specific amendments. Domain-specific models like Legal-BERT [Chalkidis et al. 2020] improved juridical modeling but remain largely restricted to span-based labeling [Araujo and Silveira 2025]. Conversely, recent advancements in "Legal Fact-Finding" demonstrate that generative architectures, particularly T5, are highly effective for transforming unstructured legal facts into structured representations [Qin and Luo 2024]. Similarly, top-performing approaches in international benchmarks (e.g., COLIEE) emphasize that designing methods based on domain characteristic observations is more impactful than model scale alone [Nguyen et al. 2024].

This study bridges the gap between statistical modeling and rule-oriented extraction by framing normative-role identification as a generative extraction task. By adopting sequence-to-sequence formulations (T5) and Semantic Role Labeling (SRL) principles, we replace traditional tagging with a text-to-structure mapping. This approach reconstructs the functional configuration of regulatory intent with greater flexibility and precision than document-centric or sequence-labeling methods.

3. Methodology

3.1. Data and Annotation

The dataset was compiled from Brazilian *Medidas Provisórias* submitted to the Federal Senate between 2016 and 2025, collected from official legislative records and randomly sampled to avoid temporal concentration. Each instance corresponds to the full text of

a single amendment (*emenda*). Pure revocations or referential edits were excluded due to insufficient semantic content. The final corpus contains 252 amendments, partitioned deterministically into training, validation, and test sets (80/10/10, seed 42).

Annotations follow a reduced, legally motivated SRL scheme capturing three roles — *Purpose*, *Recipients*, and *Restriction* — representing the minimal structure required to reconstruct normative intent, based on the study of [Humphreys et al. 2020]. Three legally trained annotators labelled the corpus, achieving Cohen's $\kappa = 0.74$, followed by adjudication.

3.2. Model Setup and Evaluation

We fine-tuned *PTT5-base* for direct text-to-structure transduction (512/256 tokens), an architecture noted for its resilience in mapping statutory language to functional roles [Qin and Luo 2024]. Optimization used AdamW (1×10^{-4}, batch 4) with early stopping and no data augmentation to isolate the effects of supervised alignment. A non-fine-tuned *T5-untuned* serves as an architectural control.

Compact general-purpose LLMs were evaluated under identical zero-shot instructions requiring the same constrained JSON schema. This deliberately asymmetric comparison characterizes the trade-off between lightweight fine-tuning and immediate usability, quantifying the "domain gap" that persists in LLMs despite their parameter scale [Deng et al. 2022]. Outputs were minimally normalized for schema validity, using beam-search decoding (width 4).

General-purpose LLMs were evaluated in a zero-shot asymmetric benchmark to quantify the "domain gap" inherent in larger scales [Deng et al. 2022]. Evaluation combined lexical metrics (ROUGE-L, METEOR), embedding similarity, and an additive expert ordinal scheme (H). The H metric (range -3 to 3) penalizes hallucinations (-1) to reflect the high stakes of legislative monitoring, where false positives are more detrimental than omissions [Nguyen et al. 2024], thus prioritizing juridical adequacy over textual overlap.

4. Results and discussion

The evidence presented in Table 1 suggests that domain adaptation may shift the operational character of the task from unconstrained generation toward a more guided semantic transformation. An aggregate analysis indicates that under conditions of higher drafting regularity, the fine-tuned T5-*legis* model appears to benefit from supervision that encodes the latent template structure of legislative amendments. This potentially leads to outputs that more consistently preserve normative relations while maintaining computational efficiency. Such improvements could be attributed to an alignment between model bias and domain regularity rather than simply to model scale. In this context, structured extraction seems to behave less like autonomous generative reasoning and more like a form of controlled canonicalisation, which may diminish the inherent advantages typically associated with larger LLMs.

As input length increases, performance variations likely reflect challenges in maintaining role localisation across extended prescriptive sequences, rather than specific deficiencies in lexical modelling. However, the schema-aware metrics in Table 1 reveal a notable disparity: while smaller general-purpose models like Llama-3.2:1b achieve higher

Table 1. Aggregate performance across models and Semantic Roles

Model	Global				Time	Purpose		Recipients		Restriction		Overall Mean	
	MET.	R-L	EM	F1	(sec.)	Sim.	H.Eval	Sim.	H.Eval	Sim.	H.Eval	Sim.	H.Eval
T5-*legis*	**0.57**	**0.60**	0.17	0.32	**1.63**	**0.77**	**0.83**	0.59	**0.74**	0.44	0.12	**0.61**	**0.56**
T5-untuned	0.54	0.50	0.28	0.36	2.29	0.68	0.00	0.23	0.00	0.30	0.00	0.40	0.00
Llama-3.2 (3b)	0.38	0.44	0.24	0.38	3.13	0.54	0.43	**0.68**	0.41	0.40	**0.30**	0.54	0.38
Gemma3:1b	0.50	0.50	0.01	0.16	1.90	0.44	0.29	0.54	0.57	**0.40**	0.21	0.46	0.35
DeepSeek-r1:1.5b	0.28	0.29	0.39	0.39	3.91	0.48	0.00	0.20	0.04	0.26	0.00	0.31	0.01
Gemma:2b	0.33	0.38	0.39	0.40	3.42	0.33	0.03	0.32	-0.02	0.33	0.09	0.32	0.03
Llama-3.2:1b	0.39	0.44	**0.44**	**0.48**	3.56	0.25	0.20	0.31	0.26	0.39	0.13	0.31	0.20
Qwen2:1.5b	0.48	0.50	0.33	0.39	3.99	0.55	0.17	0.52	0.14	0.39	0.11	0.49	0.14

Note: MET. (METEOR); R-L (ROUGE-L); EM (Exact Match); F1 (Token-level F1-Score); Sim. (Semantic Similarity); H.Eval (Human Evaluation). T5-legis refers to the domain-adapted PTT5-base model.

F1 and EM scores, this often results from verbatim text repetition. These models tend to incorporate lexical noise that inflates overlap metrics but fails to isolate the functional core of the norm.

In contrast, the decoupling between similarity-based metrics and human judgement suggests that surface correspondence is an imperfect proxy for juridical fidelity. The T5-*legis* prioritizes semantic synthesis over mere reproduction, as evidenced by its superior expert adequacy (0.83 for Purpose) despite lower lexical overlap. Furthermore, the instability noted in the *Restriction* role across architectures suggests that conditional and limiting clauses may constitute a primary source of semantic loss for most models, regardless of parameter count.

Table 2. Statistical significance and correlation with document length

Model Name	Statistical Superiority (vs. T5-*legis*)			Performance vs. Text Length		
	p-value	Cohen's d	Significance	Spearman's ρ	p-value	Interpretation
T5-*legis*	—	—	—	-0.547	0.0038	Moderate Negative
T5-untuned	0.0013	0.91	**Significant**	-0.504	0.0086	Moderate Negative
Llama-3.2 (3b)	0.0151	0.66	**Significant**	0.079	0.7025	Not Significant
Gemma3:1b	0.2578	0.27	Not Significant	0.480	0.0131	Weak Positive
DeepSeek-r1:1.5b	0.0010	1.15	**Significant**	-0.361	0.0696	Not Significant
Gemma:2b	< 0.0001	2.23	**Highly Sig.**	-0.108	0.6005	Not Significant
Llama-3.2:1b	0.0028	0.83	**Significant**	-0.148	0.4695	Not Significant
Qwen2:1.5b	0.3940	0.23	Not Significant	-0.094	0.6465	Not Significant

Wilcoxon and Cohen's d metrics are based on METEOR scores; Spearman's ρ represents the correlation between Semantic Similarity and input length. Note: Wilcoxon p-values < 0.05 indicate that T5-legis is statistically superior to the baseline. Spearman's ρ measures the monotonic relationship: negative values indicate performance degradation as text length increases. Effect size (Cohen's d) interpretation: 0.5 (Medium), 0.8 (Large), >1.2 (Very Large).

Inferential analysis (Table 2) supports this interpretation. Pairwise non-parametric tests indicate that the improvements associated with T5-*legis* are statistically distinguishable from several baselines, though not uniformly across all comparisons, reinforcing that the advantage is conditional rather than universal. Correlation analysis reveals a moderate negative association between performance and document length for encoder–decoder models, indicating sequence compression as a structural limitation. The lack of expert-level gains in models with length-insensitive automatic scores confirms that NLG metric robustness does not equate to preserving normative meaning.

The performance stability of the domain-adapted T5 model aligns with challenges documented in international benchmarks, where the complexity of statutory language often degrades the performance of general-purpose architectures [Rabelo et al. 2022]. These results therefore argue against a substitution hypothesis in which compact, domain-adapted models are simply replaced by LLMs. Instead, they delineate a task-dependent efficiency frontier where specialized encoder–decoder models are possibly better suited to domains where semantic variability is low and institutional drafting conventions are stable, while larger generative systems offer limited additional benefit for reconstructing legally structures.

More broadly, the findings caution against evaluating Legal NLP systems solely through lexical or embedding-based metrics. Since regulatory language encodes authority through functional configuration rather than paraphrasable content, reliable assessment requires coupling automatic measures with expert-informed validation capable of detecting shifts in prescriptive force. This perspective finds anchor in the studies of [Qin and Luo 2024] and [Deng et al. 2022], which argue that while LLMs excel in general reasoning, they struggle to tackle complex tasks with intricate schemas in specialized domains, where well-tuned, smaller language models provide more consistent alignment with institutional requirements.

5. Conclusion

This study evaluated whether structured legislative extraction benefits more from domain alignment than from model scale. Results indicate that supervised adaptation of a compact encoder–decoder model recovers legal relations with consistency often exceeding that of larger general-purpose LLMs.

Our findings suggest that this task is better characterized as constrained semantic transduction than open-ended generation. The resilience of the PTT5 model in mapping statutory text to a functional schema confirms the effectiveness of sequence-to-sequence architectures for legal fact-finding, as previously demonstrated in similar structured tasks [Qin and Luo 2024]. Performance was driven primarily by the alignment between training signals and institutional drafting regularities, rather than parameter count; increased scale did not systematically improve juridical adequacy. This aligns with evidence that well-tuned, smaller models often outperform general-purpose LLMs in low-resource information extraction [Deng et al. 2022]. Furthermore, our results reinforce that meticulous observation of domain-specific characteristics is a more impactful driver for reliability than model size alone [Nguyen et al. 2024].

Limitations and Future Work. The study is constrained by a small corpus. Future work will explore *Domain-Adaptive Language Engineering* (DALE) for data augmentation and pre-training to mitigate overfitting. We also intend to compare autoregressive generation against non-autoregressive frameworks (e.g., EncT5) and evaluate segmentation techniques, such as sliding windows and long-context transformers.

References

Araujo, G. and Silveira, R. (2025). Análise comparativa do bert e chatgpt no reconhecimento de entidades nomeadas do domínio jurídico. *Revista Eletrônica de Iniciação Científica em Computação*, 23:63–68.

Athan, T., Boley, H., Governatori, G., Palmirani, M., Paschke, A., and Wyner, A. (2013). Oasis legalruleml. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Law*, ICAIL '13, page 3–12. ACM.

Avgerinos Loutsaris, M., Alexopoulos, C., Maratsi, M. I., and Charalabidis, Y. (2023). Semantic interoperability for legal information: Mapping the european legislation identifier (eli) and akoma ntoso (akn) ontologies. In *Proceedings of the 16th International Conference on Theory and Practice of Electronic Governance*, ICEGOV 2023, page 41–53. ACM.

Batista, R. et al. (2021). Reconhecimento de entidades nomeadas em textos jurídicos em português. *Revista de Informática Teórica e Aplicada*.

Braz, F. A., da Silva, N. C., de Campos, T. E., Chaves, F. B. S., Ferreira, M. H. S., Inazawa, P. H., Coelho, V. H. D., Sukiennik, B. P., de Almeida, A. P. G. S., Vidal, F. B., Bezerra, D. A., Gusmao, D. B., Ziegler, G. G., Fernandes, R. V. C., Zumblick, R., and Peixoto, F. H. (2018). Document classification using a bi-lstm to unclog brazil's supreme court.

Chalkidis, I., Fergadiotis, M., Malakasiotis, P., Aletras, N., and Androutsopoulos, I. (2020). Legal-bert: The muppets straight out of law school. In *Findings of EMNLP*. Association for Computational Linguistics.

Deng, S., Ma, Y., Zhang, N., Cao, Y., and Hooi, B. (2022). Information extraction in low-resource scenarios: Survey and perspective.

Humphreys, L., Boella, G., van der Torre, L., Robaldo, L., Di Caro, L., Ghanavati, S., and Muthuri, R. (2020). Populating legal ontologies using semantic role labeling. *Artificial Intelligence and Law*, 29(2):171–211.

Nguyen, C., Nguyen, P., Tran, T., Nguyen, D., Trieu, A., Pham, T., Dang, A., and Nguyen, L.-M. (2024). Captain at coliee 2023: Efficient methods for legal information retrieval and entailment tasks.

Palmirani, M., Governatori, G., Rotolo, A., Tabet, S., Boley, H., and Paschke, A. (2011). *LegalRuleML: XML-Based Rules and Norms*, page 298–312. Springer Berlin Heidelberg.

Qin, W. and Luo, X. (2024). A legal fact-finding model based on the t5 and lexilaw large language models. In *Proceedings of the 2024 8th International Conference on Computer Science and Artificial Intelligence*, CSAI 2024, page 229–237. ACM.

Rabelo, J., Goebel, R., Kim, M.-Y., Kano, Y., Yoshioka, M., and Satoh, K. (2022). Overview and discussion of the competition on legal information extraction/entailment (coliee) 2021. *The Review of Socionetwork Strategies*, 16(1):111–133.

Vitória, J. et al. (2025). Avaliação de modelos sentence-bert para recuperação de informação legislativa. *Revista de Estudos Legislativos*.

Vitório, D., Souza, E., Dos Santos, J. A., De Carvalho, A. C. P. d. L. F., Oliveira, A. L. I., and F. da Silva, N. F. (2025). Bm25 x vila sésamo: avaliando modelos sentence-bert para recuperação de informação no cenário legislativo brasileiro. *Linguamática*, 17(1):17–33.