Contradiction Detection in Brazilian Legislative Bills: A Limited Labeled Data Study

Lucas G. L. Costa[1], **Átila M. Souza**[1], **Marcelo M. Andrade**[2], **Rodrygo L. T. Santos**[1], **Wagner Meira Jr.**[1], **Gisele L. Pappa**[1]

[1] Universidade Federal de Minas Gerais (UFMG) – Belo Horizonte, MG – Brazil
[2]Assembleia Legislativa de Minas Gerais (ALMG) – Belo Horizonte, MG – Brazil

{lucas.lage, atilamelo}@dcc.ufmg.br
marcelo.migueletto@almg.gov.br
{rodrygo, meira, glpappa}@dcc.ufmg.br

Abstract. *Identifying contradictory legislative bills is essential for ensuring consistency in Brazilian legal frameworks, yet it remains a manual and hard-to-scale task. In this work, we study automatic contradiction detection under limited labeled data. We use a dataset of 106 contradictory bill pairs curated by domain specialists and investigate two strategies: (i) supervised learning with synthetic data generation and (ii) zero-shot and few-shot classification using generative large language models (LLMs). Results show that few-shot LLMs perform strongly and outperform the supervised approach based on synthetic data. Although the small dataset limits generalization, our findings highlight the challenges of synthetic data in legal domains and the potential of LLMs for low-data scenarios.*

1. Introduction

Legislative processes involve the creation and revision of bills that propose new laws or amendments [Brasil 2025]. Ensuring consistency within this process requires identifying not only similar proposals, but also those that may lead to conflicting legal outcomes. Detecting contradictions among semantically related bills is therefore essential to avoid unintended conflicts in the legislative framework.

In the Legislative Assembly of Minas Gerais (ALMG), a Brazilian state-level legislative body, newly introduced bills are systematically linked to existing ones based on similarity, in order to prevent redundancy and promote coherence [Brasil - Minas Gerais 2025]. This practice results in a large collection of semantically related bill pairs. However, while these links capture similarity, they do not indicate whether the associated bills are contradictory, leading to a scarcity of labeled data for contradiction detection. Consequently, identifying such contradictions typically requires expert analysis of legal texts, making the process time-consuming, subjective, and difficult to scale.

In this work, we investigate the automatic detection of contradictions between legislative bills using Natural Language Processing (NLP) techniques under limited labeled data. To address this challenge, we explore two strategies: (i) supervised learning supported by synthetic data generation, and (ii) classification using generative Large Language Models (LLMs) in zero-shot and few-shot settings. We frame our study around the following research questions:

RQ1. Can synthetic data generation effectively support supervised learning for contradiction detection under limited labeled data?

RQ2. How well do LLMs perform in zero-shot and few-shot settings for identifying contradictions between legislative bills?

RQ3. How do supervised approaches based on synthetic data compare to few-shot LLMs for contradiction detection under limited labeled data?

To answer these questions, we evaluate the proposed approaches on a dataset of 106 contradictory bill pairs manually curated by domain specialists. Our findings highlight the challenges of modeling contradiction in legal texts and the potential of LLMs in settings with limited labeled data.

2. Related Work

The use of language models in brazilian governmental contexts has grown with open data initiatives that promote transparency and accessibility [Martins and Medeiros 2025, Silva et al. 2024]. In Portuguese, most approaches build on BERTimbau [Souza et al. 2020], a general-purpose model pre-trained on large-scale web corpora [Devlin et al. 2019]. Domain-specific extensions include GovBERT-BR [Silva et al. 2025], which continues pre-training on governmental texts, and JurisBERT [Viegas et al. 2023], trained from scratch on Brazilian legal data. Similarly, RoBERTaLexPT [Garcia et al. 2024] focuses on legal corpora.

Recent work on contradiction detection in legal texts has primarily relied on supervised BERT-based models, often framing the task as sentence-pair classification and using data augmentation or self-training to mitigate limited labeled data [Navastara et al. 2025, Zaiton et al. 2026]. In this context, we investigate the role of synthetic data for supervised learning and evaluate LLMs in zero-shot and few-shot settings under limited labeled data.

3. Data

Our study is based on legislative data from the Legislative Assembly of Minas Gerais (ALMG). The corpus comprises 49,868 legislative bills, including 3,754 pairs previously identified as similar according to institutional procedures [1]. However, these links do not capture whether the associated bills are contradictory, resulting in a scarcity of labeled data for contradiction detection.

To address this limitation, we rely on a curated dataset of 106 pairs of contradictory legislative bills [1] , identified by domain specialists based on legislative analysis documents. Additionally, we conducted an analysis to categorize each pair of contradictory bills by their central theme. This categorization supports both the design of experiments and interpretation of results. The following presents the distribution of categories across the 106 contradictory bill pairs, along with a brief description of each category:

Rodeo (48 pairs - 45.3%). bills focused on animal protection in opposition to bills that recognize rodeos as cultural heritage.

COVID-19 Vaccination (12 pairs - 11.3%). bills that restrict access of unvaccinated individuals to public spaces versus bills that prohibit such restrictions.

[1]Data available at: `https://github.com/LucasLage/legislative-bill-contradiction`

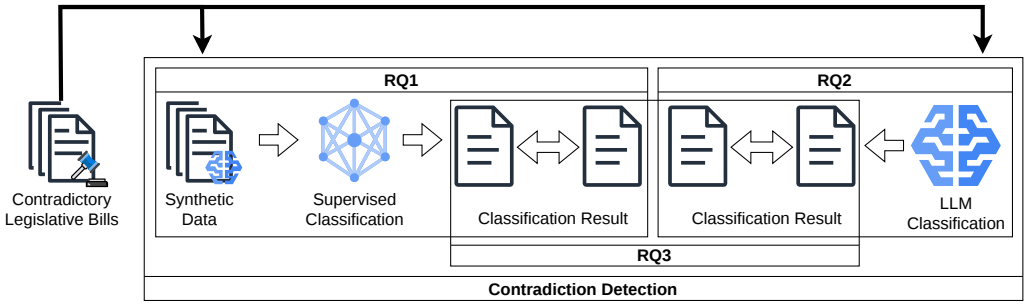

Figure 1. Overview of the proposed methodology for contradiction detection between legislative bills under limited labeled data.

Apology of Crime (6 pairs - 5.7%). bills that prohibit funding for artists allegedly promoting crime versus bills that restrict subjective censorship.

Road Concessions (3 pairs - 2.8%). bills that mandate public disclosure of information versus bills that delegate such decisions to concession agreements.

Funk (4 pairs - 3.8%). bills that recognize funk as a cultural expression and promote its inclusion in schools versus bills that restrict inappropriate events in educational settings.

Gender (7 pairs - 6.6%). bills that define gender based on biological criteria versus bills that recognize self-identified gender and the use of social names.

Others (26 pairs - 24.5%). pairs whose topics are unique within the dataset.

4. Methodology

We adopt two strategies to study contradiction detection under limited labeled data, illustrated in Figure 1. The first addresses RQ1 by expanding the training set through synthetic data generation (Section 4.1), enabling supervised fine-tuning of a BERT-based model (Section 4.2). The second addresses RQ2 by leveraging (LLMs) in zero-shot and few-shot settings (Section 4.3). These approaches are later compared to address RQ3.

4.1. Synthetic Data Generation

To mitigate the limited size of the labeled dataset, we generated synthetic examples of contradictory legislative bills. Each synthetic pair consists of a real legislative bill and a corresponding contradictory version generated by a Large Language Model.

From the original set of 49,868 bills in the ALMG dataset, a filtering step was applied to select suitable candidates for generation. Bills with excessively long texts were removed to reduce the likelihood of hallucinations during generation, as well as less informative bills such as those establishing commemorative dates, declaring public utility, specifying property donations, or assigning names. This filtering process resulted in a subset of 17,077 bills.

The synthetic generation was performed using the Mistral-Small-24B-Instruct-2501 model [Mistral AI 2025], quantized to 4-bit precision (NF4) and running inference in bfloat16. Text generation used stochastic decoding with a temperature of 0.8 and top-p sampling of 0.95. Generation was conducted in a few-shot setting, where 10 real examples of contradictory bill pairs (drawn from the 106 manually curated pairs) were included in

the prompt as guidance. In addition to these examples, the prompt contained the original bill, and the model was instructed to generate a new bill that contradicts the original while preserving its structure, style, and thematic context. This process resulted in 17,077 synthetic contradictory bills [2], each paired with a real bill, enabling the construction of a larger dataset for supervised training.

4.2. Supervised BERT-based Classification

We fine-tune JurisBERT [Viegas et al. 2023] for the binary classification of contradictions between legislative bills. JurisBERT was selected because, among the pre-trained BERT-based models described in the literature, it was trained from scratch on Brazilian legal corpora, making it particularly suitable for Brazilian legal language.

The task is formulated as a sentence-pair classification problem, in which the model takes a pair of bills and predicts whether they are contradictory or not. The model is fine-tuned on a dataset of synthetic contradictory pairs, as described in Section 4.1, complemented with non-contradictory examples. Leveraging domain-specific pre-training, the model is expected to capture legal language patterns relevant to identifying inconsistencies between bills. This setup allows us to evaluate whether synthetic data can support supervised learning under limited labeled data.

4.3. LLM-based Classification

As an alternative to supervised learning, we evaluated the use of generative Large Language Models (LLMs) for direct classification without task-specific training. Two scenarios were considered to assess their effectiveness under limited labeled data: (i) Zero-shot: the model receives only task instructions. (ii) Few-shot: the model receives a small set of labeled examples before the task instructions.

We evaluated three models: (i) Mistral-Small-24B-Instruct-2501 model [Mistral AI 2025] in the same quantization parameters as explainned in Section 4.1, (ii) GPT-5 nano [OpenAI 2025b], and (iii) GPT-5 mini [OpenAI 2025a]. All models were evaluated with temperature set to zero. The GPT-5 models used top-p equal to 1 and a fixed random seed of 42, while the Mistral model employed deterministic decoding.

5. Experimental Evaluation

In this section, we evaluate four models for contradiction detection between legislative bills. Section 5.1 describes the supervised fine-tuning procedure, Section 5.2 outlines the evaluation setup, Section 5.3 presents the results.

5.1. Supervised Fine-Tuning

The dataset [2] for supervised fine-tuning on JurisBERT is composed of three types of pairs: (i) 17,077 synthetic contradictory pairs; (ii) 3,695 semantically similar but non-contradictory pairs; (iii) 7,390 randomly selected non-related pairs. This composition aims to expose the model to both hard negatives (similar but non-contradictory) and easy negatives (unrelated pairs), improving its ability to distinguish true contradictions from other types of relationships. The dataset was split into training (70%), validation (10%), and test (20%) sets.

[2]Data available at: `https://github.com/LucasLage/legislative-bill-contradiction`

Fine-tuning was executed for five epochs, employed the AdamW optimizer with a learning rate of 5e-5, $\beta_1 = 0.9$, $\beta_2 = 0.999$, weight decay of 0.01, dropout probability of 0.1, and a warm-up ratio of 0.1. All other settings used the default parameters of the Hugging Face Transformers library. [3]

5.2. Evaluation Setup

The evaluation is designed to enable a direct comparison between supervised and LLM-based approaches under limited labeled data, addressing the research questions.

Four models were evaluated, one BERT-based model (JurisBERT) fine-tuned in a supervised manner on synthetic and real dataset. And three generative Large Language Models (LLMs), in zero-shot and few-shot settings: Mistral-Small-24B (quantized), GPT-5 nano, and GPT-5 mini. All models were evaluated on the same types of data pairs, including real contradictory bill pairs, synthetic contradictory pairs, semantically similar but non-contradictory pairs, and randomly selected non-related pairs.

To ensure a rigorous evaluation and mitigate the impact of sample variability, we adopted four experimental splits [4] derived from the 20% test partition of the supervised fine-tuning dataset together with the 106 real contradictory bill pairs. Each split contains a test set and a set of few-shot examples, both varying across splits. Across the four splits, all real contradictory pairs are used, with 100 allocated to the test set and the remaining 6 used as few-shot examples.

Each test set consists of (i) 200 contradictory pairs, including 100 real and 100 synthetic examples, and (ii) 200 non-contradictory pairs, comprising 100 semantically similar pairs and 100 randomly selected unrelated pairs. The composition of few-shot examples in each split was designed to ensure thematic diversity. Each split includes 12 example pairs, comprising 6 real contradictory pairs and 6 non-contradictory pairs. The six real contradictory pairs used as few-shot examples in each split are drawn from the following categories:

Split 1. all examples belong to the ”Others” category.

Split 2. Rodeo, Gender, and four from ”Others”.

Split 3. Rodeo, COVID-19 Vaccination, Apology of Crime, and three from ”Others”.

Split 4. Rodeo, Funk, Road Concessions, and three from ”Others”.

As we address a classification task, we evaluate model performance using Accuracy and Macro F1-score. Models were evaluated on each split, and results are reported as the mean and standard deviation across the four splits.

5.3. Results

We report results with respect to the research questions defined in Section 1. Across all four splits, all models achieved Accuracy and Macro F1-score equal to 1.0 on both synthetic contradictory pairs and randomly selected non-contradictory pairs. These results indicate that such cases are trivial, where contradiction is explicit and easily identifiable.

[3] `https://huggingface.co/docs/transformers/`

[4] Data available at: `https://github.com/LucasLage/legislative-bill-contradiction`

Table 1. Results across all models and datasets.

	All data		Remove trivials	
Model	**Accuracy**	**Macro F1**	**Accuracy**	**Macro F1**
JurisBERT-ft	0.858 ± 0.002	0.846 ± 0.003	0.716 ± 0.004	0.691 ± 0.005
Mistral-Small Zero Shot	0.871 ± 0.005	0.870 ± 0.006	0.742 ± 0.011	0.739 ± 0.011
GPT-5 nano Zero Shot	0.856 ± 0.003	0.856 ± 0.003	0.719 ± 0.006	0.712 ± 0.006
GPT-5 mini Zero Shot	0.845 ± 0.004	0.843 ± 0.004	0.690 ± 0.008	0.685 ± 0.009
Mistral-Small Few Shot	0.893 ± 0.003	0.893 ± 0.003	0.786 ± 0.005	0.786 ± 0.005
GPT-5 nano Few Shot	0.868 ± 0.008	0.866 ± 0.008	0.735 ± 0.015	0.731 ± 0.017
GPT-5 mini Few Shot	**0.895 ± 0.028**	**0.895 ± 0.028**	**0.790 ± 0.055**	**0.789 ± 0.056**

Table 2. Accuracy per class for the real contradictory pairs and the semantically similar but non-contradictory pairs.

	Accuracy	
Model	**Real contradictory**	**Similar non-contradictory**
JurisBERT-ft	0.432 ± 0.008	0.998 ± 0.000
Mistral-Small Zero Shot	0.625 ± 0.009	0.860 ± 0.016
GPT-5 nano Zero Shot	0.570 ± 0.007	0.867 ± 0.019
GPT-5 mini Zero Shot	0.570 ± 0.020	0.810 ± 0.007
Mistral-Small Few Shot	0.810 ± 0.055	0.762 ± 0.052
GPT-5 nano Few Shot	0.618 ± 0.035	0.852 ± 0.008
GPT-5 mini Few Shot	0.755 ± 0.096	0.825 ± 0.018

As a result, evaluations including these examples overestimate model performance, reinforcing the importance of focusing on more challenging subsets.

Table 1 reports results under two evaluation settings: (i) the full dataset and (ii) a challenging subset excluding trivial examples. The latter retains only real contradictory pairs and semantically similar non-contradictory pairs, preserving class balance while reflecting the target task. Table 2 reports class-wise accuracy on this non-trivial subset.

RQ1: Synthetic data for supervised learning. The JurisBERT model was fine-tuned on a combination of synthetic contradictory pairs and non-contradictory data, with 20% reserved for testing. Under this setting, the model achieved Accuracy and Macro F1-score above 0.98. However, when evaluated on the more challenging non-trivial subset (Table 1), performance drops substantially.

Further analysis (Table 2) reveals a strong bias: the model achieves near-perfect accuracy on semantically similar non-contradictory pairs, but performs poorly on real contradictory pairs. This indicates that the model primarily captures semantic similarity rather than contradiction patterns.

These results highlight a limitation of the synthetic data, which tends to produce overly trivial examples whose learned patterns fail to generalize to real contradiction detection. Overall, the findings suggest that synthetic data alone is insufficient to effectively support supervised learning in this setting.

RQ2: LLM performance in zero-shot and few-shot settings. In the zero-shot setting, Mistral-Small-24B achieves the best performance, followed by GPT-5 nano, while GPT-5 mini performs worst. In the few-shot setting, all LLMs improve, confirming the effectiveness of minimal supervision. GPT-5 mini shows the largest gains and achieves the best

overall performance, albeit with higher variance across splits.

Class-wise analysis (Table 2) shows that zero-shot LLMs tend to favor non-contradictory pairs, while few-shot prompting shifts performance toward improved detection of real contradictions. Notably, GPT-5 mini is the only model that improves performance in both classes, indicating a more effective use of the few-shot examples. These results demonstrate that LLMs, particularly in the few-shot setting, are effective for contradiction detection under limited labeled data.

RQ3: Comparison between supervised and LLM-based approaches. Table 1 shows that all models achieve substantially higher performance on the full dataset, where trivial examples are present. When restricting the evaluation to the non-trivial subset, performance decreases across all models, providing a more realistic assessment.

Under this setting, few-shot LLMs consistently outperform the supervised JurisBERT model. While the supervised model benefits from large-scale synthetic data during training, it fails to generalize to real contradictory cases. In contrast, LLMs, particularly in the few-shot setting, exhibit more balanced performance across classes and better capture real contradiction patterns. Mistral-Small achieves the highest accuracy on real contradictory pairs, while GPT-5 mini achieves the best overall performance.

Limitations. Despite these findings, several limitations must be acknowledged. The evaluation dataset is relatively small, constraining the assessment of generalization, and exhibits thematic imbalance, which may introduce bias. Additionally, synthetic data proved to be trivial to classify, limiting its usefulness for robust evaluation.

6. Conclusion

This work investigated contradiction detection between legislative bills under limited labeled data, focusing on the role of synthetic data and large language models (LLMs).

Our findings show that synthetic data alone is insufficient to effectively support supervised learning for this task. Although it enables large-scale training, it tends to produce trivial examples that fail to capture the complexity of real-world contradictions, limiting generalization. In contrast, LLMs demonstrate strong performance, particularly in the few-shot setting, where even a small number of examples substantially improves their ability to identify real contradictions. These results highlight the potential of LLMs for contradiction detection in legal text.

A key limitation of this study is the small size of the evaluation dataset, which constrains the assessment of generalization, as well as its thematic imbalance. Future work includes exploring hybrid approaches that combine retrieval and reasoning capabilities, as well as investigating more effective strategies for generating non-trivial synthetic data.

Acknowledgments. This work was supported by Legislative Assembly of Minas Gerais (*Assembleia Legislativa de Minas Gerais*, ALMG) and by CNPq, CAPES, FAPEMIG.

Declaration of Generative AI Use. Generative AI tools were used solely for grammatical revision and to improve text fluency.

References

Brasil (2025). Regimento interno da Câmara dos Deputados. `https:`

//www2.camara.leg.br/atividade-legislativa/legislacao/regimento-interno-da-camara-dos-deputados. Accessed: 2026-05-01.

Brasil - Minas Gerais (2025). Regimento interno da Assembleia Legislativa do Estado de Minas Gerais. https://www.almg.gov.br/atividade-parlamentar/leis/legislacao-mineira/lei/texto/?tipo=RAL&num=5176&ano=1997&comp=&cons=1. Accessed: 2026-05-01.

Devlin, J., Chang, M., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT*, pages 4171–4186. Association for Computational Linguistics.

Garcia, E., Silva, N., Siqueira, F., Gomes, J., Albuquerque, H. O., Souza, E., Lima, E., and de Carvalho, A. (2024). RoBERTaLexPT: A legal RoBERTa model pretrained with deduplication for Portuguese. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 1*, pages 374–383, Santiago de Compostela, Galicia/Spain. Association for Computational Lingustics.

Martins, M. and Medeiros, C. (2025). From anvisa leaflets to extended interoperability with global health databases: some pitfalls and success stories. In *Anais do XIX Brazilian e-Science Workshop*, pages 9–16, Porto Alegre, RS, Brasil. SBC.

Mistral AI (2025). Mistral-small-24b-instruct-2501. https://huggingface.co/mistralai/Mistral-Small-24B-Instruct-2501. Accessed: 2026-05-01.

Navastara, D. A., Abdillah, S., Benito, D., Adillion, I. G., and Purwitasari, D. (2025). Document matching for contradiction detection in low-resource legislative texts with self-training and augmentation using transformer model. *JANAPATI*, 14(2):321–335.

OpenAI (2025a). Gpt-5 mini. https://developers.openai.com/api/docs/models/gpt-5-mini. Accessed: 2026-05-01.

OpenAI (2025b). Gpt-5 nano. https://developers.openai.com/api/docs/models/gpt-5-nano. Accessed: 2026-05-01.

Silva, M., Oliveira, G., Costa, L., and Pappa, G. (2024). Evaluating domain-adapted language models for governmental text classification tasks in Portuguese. In *Anais do XXXIX SBBD*, pages 247–259, Porto Alegre, RS, Brasil. SBC.

Silva, M. O., Oliveira, G. P., Costa, L. G. L., and Pappa, G. L. (2025). GovBERT-BR: A BERT-based language model for Brazilian Portuguese governmental data. In *Intelligent Systems. BRACIS 2024*, pages 19–32, Cham. Springer Nature Switzerland.

Souza, F., Nogueira, R., and Lotufo, R. (2020). BERTimbau: pretrained BERT models for Brazilian Portuguese. In *9th BRACIS, Rio Grande do Sul, Brazil, October 20-23*.

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

Zaiton, H., Alansary, S., Sarwat, N., and Hussein, I. (2026). Automatic detection of contradictions in legal texts: A computational linguistic approach. *Procedia Computer Science*, 275:503–512. 7th ACling.