Performance Analysis of LLMs for Abstractive Summarization of Brazilian Legislative Documents

Danilo C. G. de Lucena[a*], *Ellen* Souza[b], *Hidelberg O.* Albuquerque[c], *Nádia* Félix[d], *Adriano L.I.* Oliveira[e], *André C.P.L.F.* de Carvalho[f]

[a] Centro de Informática, Federal University of Pernambuco, Recife, Brazil, dcgl@cin.ufpe.br.

[b]MiningBR Research Group, Federal Rural University of Pernambuco, Recife, Brazil, ellen.ramos@ufrpe.br.

[c]Centro de Informática, Federal University of Pernambuco, Recife, Brazil, hidelberg.albuquerque@ufrpe.br.

[d]Institute of Informatics, Federal University of Goiás, Goiás, Brazil, nadia.felix@ufg.br.

[e]Centro de Informática, Federal University of Pernambuco, Recife, Brazil, alio@cin.ufpe.br.

[f]Institute of Mathematics and Computer Sciences, University of São Paulo, São Paulo, Brazil, andre@icmc.usp.br.

Submitted: 31 January 2025, Revised: 26 March 2025, Accepted: 21 April 2025, Published: 20 May 2025

Abstract. Legislative documents present substantial obstacles to summarization due to their complex argument structures and specialized terminology. This research investigates the application of Large Language Models (LLMs) in summarizing Brazilian legislative proposals from the Chamber of Deputies, examining a dataset of over 56 thousand texts from 2013 to 2023. The paper explores three main summarization methodologies: extractive, abstractive, and hybrid, with an emphasis on abstractive summarization using LLMs. The performance of the LLM LLAMA2-13b is assessed using metrics such as ROUGE, BLEU, METEOR, BERTScore, and BERTopic, compared against reference summaries. The results show that LLMs can generate coherent and informative summaries, with positive evaluation metric results. Notably, the study reveals that traditional summary evaluation metrics may not be adequate for evaluating LLMs in summarization tasks. On the other hand, metrics based on pre-trained models like BERT provide a more effective evaluation of this innovative automatic summarization approach.

Keywords. large language models, summarization, legislative proposals.
Research paper, DOI: https://doi.org/10.59490/dgo.2025.969

1. Introduction

In the domain of Natural Language Processing (NLP), text summarization is the process of automatically generating a concise summary that retains the most relevant information from a longer text document, preserving the core or key ideas from the original document (Neto et al., 2002). The academic literature in the NLP field identifies three main types of text summarization methods: (i) extractive, (ii) abstractive, and (iii) hybrid summarization. Extractive summarization selects the most relevant sentences, phrases, or passages from the original text to build the summary without modifying the extracted content. Abstractive summarization employs state-of-the-art natural language processing techniques to interpret the text at a semantic level and generate new content and sentences for the summary. These generated summaries may be fully rephrased or completely rewritten, rather than directly extracting text spans. Hybrid summarization combines techniques that merge extractive and abstractive summarization, typically extracting key text spans from the original text and then rewriting or enhancing them for conciseness and readability (Abualigah et al., 2020; Egan et al., 2022; El-Kassas et al., 2021; Ermakova et al., 2019; Liu et al., 2022; Tas & Kiyani, 2007).

This study investigates the application of Large Language Models (LLMs) for abstractive summarization of legislative proposals from Brazil's Chamber of Deputies over the period from 2013 to 2023. The study utilizes a corpus (Souza, Vitório, et al., 2021) containing over 56 thousand legislative proposal texts, complete with textual data, metadata, and pre-generated reference summaries. Within the scope of LLMs, the LLAMA2 (refer to Section 4) model was chosen to produce an automated summary for each legislative proposal text in the dataset. The proposed work addresses the task of summarizing legal documents, a recurrent task in the domain of Legal Natural Language Processing (Jain et al., 2021). This task aims to automatically generate concise overviews that capture the key information from lengthy and complex legal texts, such as contracts, case law, regulatory documents, and government legislative documents. Legal summarization, as a specific field of knowledge, poses unique challenges compared to other domains due to the specialized terminology, complex argument structure, and the need to preserve the legal validity of summaries.

2. Related Works

2.1. Summarization in Legal Domain

In this section, we describe studies that are important for understanding the importance of summarization in legal documents. Jain et al. provide a comprehensive state-of-the-art review that categorizes summarization techniques along two dimensions: domain-specificity (independent vs. specific) and approach type (extractive vs. abstractive). Their analysis predominantly covers extractive methods while offering limited discussion of abstractive summarization, despite the latter's greater complexity and potential for legal applications (Jain et al., 2021). Lucke et al. employed a technology-neutral methodology and multi-criteria analysis to identify several potential AI applications within parliamentary legislative processes. Their research prioritized these applications by relevance and significance, highlighting four key areas: intelligent scrutiny of legislative proposals, machine-readable legislation conversion, smart law implementation, and AI-enhanced parliamentary transparency (von Lucke et al., 2022).

Galgani et al. present a hybrid approach for legal text summarization focused on catchphrase creation for case reports. Their evaluation compares extraction methods (rule-based, citation-based, and general-purpose) using Rouge scores. The rule-based system achieved higher precision, while combining it with the citation-based method improved recall without sacrificing precision (Galgani et al., 2012). In the deep learning domain, Anand & Wagh implemented a semi-supervised approach for legal document extractive summarization using neural networks. They generated labeled data from reference summaries and evaluated quality through a set of metrics against human-generated headnotes. Their model demonstrated superior performance, validating sentence transformation effectiveness for labeled data creation (Anand & Wagh, 2022).

In the field of natural language processing for Portuguese legal documents, we refer to specific initiatives in analyzing legislative documents for the Chamber of Deputies of Brazil (Albuquerque et al., 2022; Souza, Moriyama, et al., 2021; Souza, Vitório, et al., 2021; Vitório et al., 2022, 2023).

2.2. Summarization with LLMs

In related literature, we highlight conversational Large Language Models (LLMs) — models specifically fine-tuned for dialogue-based interactions — as an emerging approach for automated text summarization. The application of these conversational LLMs to legal text summarization remains an evolving field. We believe that the studies discussed below offer valuable methodological insights for future research in this domain. Throughout our experimental process, we confirmed that the approaches examined in this section demonstrate applicability to legal text summarization scenarios.

Luo et al. (Luo et al., 2023) examined ChatGPT's capability to evaluate factual inconsistencies in text summarization using zero-shot learning, a technique where a language model performs a task without any prior examples or task-specific training, relying solely on its pre-trained knowledge and the instructions provided in the prompt. Their study employed various prompts for entailment inference, summary ranking, and consistency rating tasks, comparing ChatGPT's performance against established evaluation metrics across multiple datasets. While their results demonstrated ChatGPT's potential, they identified limitations in handling abstractive summaries and prompt alignment, suggesting future research directions in prompt engineering and alignment optimization.

In related work, Wang et al. (Wang et al., 2023) investigated ChatGPT as an evaluation metric for Natural Language Generation (NLG). Their experiments spanned five NLG meta-evaluation datasets, comparing ChatGPT with existing automatic metrics. Their findings revealed ChatGPT's strong correlation with human judgments, its sensitivity to prompt formulation, and the significant influence of meta-evaluation dataset creation methods on NLG metric effectiveness.

In the field of Generative Pre-trained Transformers (GPT) (Brown et al., 2020), Goyal et al. (Goyal et al., 2022) examined the effectiveness of prompt-based models for text summarization, using GPT-3 as their primary case study. Their research compared GPT-3 against both fine-tuned models and a multi-task instruction-tuned model, employing parallel human and automatic evaluations across diverse summarization tasks. Their findings revealed that human annotators consistently preferred GPT-3-generated summaries, which demonstrated notable flexibility in handling tasks with varying constraints. However, they identified a significant evaluation gap: current automatic metrics failed to accurately capture the quality and diversity of GPT-3 summaries, frequently assigning them lower scores despite their superior human ratings.

3. Legislative Proposal Text Dataset

This section presents a quantitative and qualitative analysis of the corpus of legislative documents of the Brazilian Chamber of Deputies (Souza, Vitório, et al., 2021). The corpus comprises several types of legislative proposals, mainly: Bill or Law Project (*"Projeto de Lei" - PL*), Complementary Bill or Law Project (*"Projeto de Lei Complementar" - PLC*), and Constitutional Amendment Proposal (*"Proposta de Emenda Constitucional" - PEC*), totaling 56,391 documents. Each document in the corpus has attributes used in this experiment: proposal identification (author's name, proposal date), original text of the proposal, and manually generated reference summaries. For the experiment conducted, the summary present in the dataset is considered the reference summary, which will be compared with the summary generated by the LLM. Other attributes were not considered in the experiment. In the experiment, the reference summaries were obtained through a process of manual summarization, carried out by the creators of the original dataset. For examples of the structure of a legislative proposal document, we refer to the document site of the Brazilian Chamber of Deputies[1].

Table 1 presents a numerical description of the corpus, detailed by year, that shows the volume of proposals generated during the period analyzed. The textual data, spanning from 2013 to 2023, show the overall variation in the number of documents per year, starting from 5,394 proposals in 2013, peaking at 6,724 in 2021, then decreasing to 4,657 in 2022, and further to 2,389 in 2023. The average text length of analyzed documents also saw significant changes, increasing from 606 words in 2013 to 782 words in 2019, followed by an increase to 1,057 words in 2023.

Throughout the analyzed period, the average length of the reference summaries remained stable, ranging from 31 to 35 words. In the dataset, all reference summaries were written manually. In contrast, the average length of summaries generated by the LLM model varied, starting at 96 words in 2013 and reaching 127 words in 2023. This discrepancy in summary lengths is attributed to the fact that creating reference summaries is more labor-intensive and time-consuming, as they depend on the availability of individuals to perform manual summarization, while the automated generation of summaries using LLMs requires mainly computational resources. This shows the model's capability to understand the text semantically in greater depth and provide longer and more elaborate summaries.

4. Experiments

4.1. Documents Preprocessing

For the preprocessing stage of the proposal texts and summaries, it was necessary to develop a pipeline for text cleaning to make them suitable for summarization tasks. In summary, the steps adopted in the study were: (i) analyzing the text to remove punctuation, line breaks, and unnecessary spacing; (ii) removing meaningless words and special characters; and (iii) sentence segmentation and tokenizing words from reference summaries for use with evaluation metrics.

[1] https://www2.camara.leg.br/a-camara/programas-institucionais/experiencias-presenciais/parlamentojovem/sou-estudante/material-de-apoio-para-estudantes/modelo-de-projeto-de-lei

Tab. 1 – Annual statistics of Brazilian legislative proposals (2013-2023). The table presents yearly document counts, average proposal length (in words), and summary length comparison between human-generated reference summaries and LLM-generated summaries. Note the consistent discrepancy between reference and LLM summary lengths across all years.

Documents			Avg. Summary Length (words)	
Year	**Document Count**	**Avg. Length**	**Reference**	**LLM**
2023	2,389	1,057.0	35.0	127.0
2022	4,657	928.5	35.1	110.5
2021	6,724	899.7	33.3	109.5
2020	6,666	986.2	34.7	117.6
2019	8,257	782.6	31.2	106.3
2018	3,116	835.2	33.3	103.6
2017	5,336	738.6	31.2	102.8
2016	4,090	804.8	30.8	103.5
2015	7,156	768.4	31.7	105.3
2014	2,606	773.1	31.7	103.2
2013	5,394	606.2	31.3	96.9

4.2. LLM and Prompt Engineering

The LLAMA2-13b model[2], a cutting-edge language model by Meta, was selected for the experiments (Touvron et al., 2023). As a member of the LLAMA2 family, it is one of several generative text models pre-trained and fine-tuned with parameters ranging from 7 billion to 70 billion.

The LLAMA2-13b model was chosen due to its optimal balance of cost and performance, as these models are known for their computational intensity and expense. This auto-regressive language model utilizes an optimized transformer architecture and has been fine-tuned through supervised fine-tuning and reinforcement learning with human feedback (RLHF) (Christiano et al., 2017), ensuring alignment with human preferences for helpfulness and safety. For the experiments, the LLAMA2-13b model was deployed in its dialogue-optimized format on the Huggingface platform [3].

Figure 1 illustrates the text summarization process using the LLAMA2-13b model. The summarization flow begins with two initial inputs: the original "Document" to be summarized and a "Reference Summary" that serves as a quality benchmark for the generated summary. The "Text Preprocessing" stage represents an intermediate step where the original document undergoes processing to eliminate unnecessary information, normalize the text, and prepare it for the summarization model. At the "Large Language Model" stage, a zero-shot prompt is provided to summarize the text. The "Input" and "Output" components indicate the language model's processing flow: the "Preprocessed Text" resulting from the text preprocessing step serves as input, while the "Summary" represents the final summarized text output.

For generating summaries with the LLAMA2-13b model, we employed specific parameters empirically determined through iterative experimentation until summary generation achieved reasonable adequacy: (i) $temperature = 0.01$, the lowest possible value to prevent model hallucination (Rawte et al., 2023); (ii) $truncate = 4,096$, to avoid processing texts with character counts exceeding this established threshold; and (iii) $max\_new\_tokens = 500$ as the maximum number of tokens to be generated in the summary.

5. Main Results

This section presents the outcomes of the metrics outlined in Section 1, providing both quantitative and qualitative evaluations divided into two assessment categories. In the first stage, we analyze the ROUGE, BLEU, and METEOR metrics that evaluate results at a syntactic level, analyzing words or sentences. In the second stage, we employ BERTScore and BERTopic metrics that leverage pretrained language models based on the BERT architecture for semantic-level text analysis, utilizing multilingual BERT models pretrained for Portuguese language analysis.

[2] https://ai.meta.com/llama/

[3] https://huggingface.co/blog/llama2

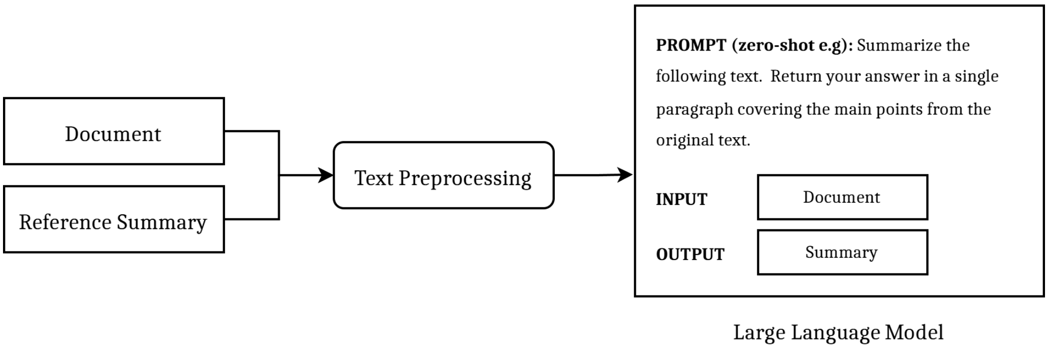
Document
Reference Summary
Text Preprocessing
PROMPT (zero-shot e.g): Summarize the following text. Return your answer in a single paragraph covering the main points from the original text.
INPUT
Document
OUTPUT
Summary
Large Language Model

Fig. 1 – Text summarization pipeline architecture using a Large Language Model. The flowchart illustrates the complete process: (1) initial inputs consisting of the original document and its reference summary; (2) text preprocessing phase that normalizes and prepares the document for summarization; (3) the Large Language Model component featuring a zero-shot prompt, with clearly defined input (preprocessed document) and output (generated summary) pathways. The reference summary serves as an evaluation benchmark though this comparison is not explicitly shown in the diagram. This architecture was implemented using LLAMA2 as the underlying language model.

5.1. Structural Level Evaluation

This proposal utilized a set of ROUGE metrics (Lin, 2004) to evaluate lexical overlap between LLM-generated and reference summaries. Specifically, ROUGE-1 and ROUGE-2 measure unigram and bigram overlap respectively, while ROUGE-L assesses the longest common subsequence. An n-gram is a contiguous sequence of n items (typically words or characters) from a text, used as a basic unit of analysis in natural language processing to evaluate textual overlap in metrics such as ROUGE-1 and ROUGE-2 in summary quality evaluation.

The ROUGE metrics analysis reveals a consistent downward trend in LLM summarization performance over the 2013-2023 period, with all three measures (Unigram, Bigram, and L-score) showing peak values in 2013 (0.399, 0.319, and 0.346 respectively) and lowest values in 2023 (0.336, 0.260, and 0.290), representing a decline. All scores remained relatively low (below 0.4), which may reflect the considerable length disparity between reference summaries (31-35 words) and LLM-generated summaries (96-127 words) rather than actual quality deficiencies. The parallel patterns across all three metrics suggest that lexical overlap at different granularities is similarly affected by underlying factors such as increasing document complexity (606 words in 2013 to 1,057 words in 2023). These findings underscore the limitations of using lexical-based evaluation metrics for abstractive summarization tasks, particularly as legislative language evolves over time.

The *Bilingual Evaluation Understudy* (BLEU) (Lin & Och, 2004; Papineni et al., 2002) is a metric that was initially developed for the purpose of evaluating machine translation systems by comparing candidate translations to reference translations. Furthermore, it has gained relevance in other natural language processing tasks like assessing text summarization systems.

The BLEU score analysis for summarization reveals a consistent temporal decline from 2013 (0.173) to 2023 (0.140), representing a performance decrease. Despite minor fluctuations in 2018 (0.158) and 2021 (0.164), all scores remain notably low (below 0.18), with a year average of 0.149. This pattern parallels the increasing document length (606.2 to 1,057 words) over the same period, suggesting a potential correlation between source complexity and summarization difficulty. The substantial length disparity between reference (31-35 words) and LLM-generated summaries (96-127 words) further complicates interpretation, as BLEU primarily measures lexical precision without capturing semantic equivalence. These findings align with the study's position that traditional n-gram overlap metrics have inherent limitations for evaluating domain-specific abstractive summarization. LLMs may be predisposed to producing more verbose summaries than what human preferences might dictate. This observation suggests a promising avenue for future research, particularly in the realm of prompt engineering for summarization tasks. Notably, there exists a strong correlation (*R=0.96*) between the lengths of summaries generated by LLAMA2 and those of manual reference summaries. As the length of reference summaries increases, the model tends to generate proportionally longer summaries.

Tab. 2 – The table displays the evaluation results of the ROUGE, BLEU, and METEOR metrics. The ”Year” column specifies the year of the text subset. The ”Unigram score” and ”Bigram score” columns represent the ROUGE metric values, calculated using unigram (1-gram) and bigram (2-gram) respectively, by comparing the generated summary with the reference summary. The ”L-score” column shows the ROUGE metric value based on the Longest Common Subsequence, again comparing the generated summary with the reference summary. The ”BLEU score” and ”METEOR score” columns present the results for these two metrics.

Year	**ROUGE**			**BLEU-score**	**METEOR-score**
	Unigram-score	**Bigram-score**	**L-score**		
2023	0.336	0.260	0.290	0.140	0.459
2022	0.369	0.283	0.313	0.150	0.475
2021	0.384	0.305	0.330	0.164	0.500
2020	0.352	0.271	0.305	0.150	0.463
2019	0.360	0.279	0.311	0.150	0.473
2018	0.379	0.292	0.326	0.158	0.479
2017	0.360	0.275	0.304	0.145	0.466
2016	0.352	0.265	0.300	0.145	0.459
2015	0.348	0.258	0.293	0.142	0.452
2014	0.371	0.284	0.317	0.152	0.475
2013	0.399	0.319	0.346	0.173	0.511

METEOR (*Metric for Evaluation of Translation with Explicit Ordering*) (Banerjee & Lavie, 2005) is a metric initially developed for machine translation evaluation that has since been widely adopted for summarization assessment. METEOR evaluates similarity between generated and reference summaries by conducting recall analysis at multiple linguistic levels: exact matching, stem matching, synonym alignment, and paraphrase correspondence. In its scoring algorithm, METEOR assigns higher weights to exact token matches than to stems and synonyms, prioritizing precision in word choice over semantic approximations.

The METEOR evaluation shows scores ranging from 0.452 (2015) to 0.511 (2013), with a mean of 0.474 across the. Unlike the consistently declining pattern in ROUGE and BLEU metrics, METEOR exhibits more fluctuation, with a secondary peak in 2021 (0.500) before dropping to 0.459 in 2023. METEOR scores substantially outperform BLEU scores across all years, reflecting its design to capture semantic similarity through exact matches, stems, synonyms, and paraphrases rather than strict n-gram correspondence. The moderate absolute performance (scores 0.5) indicates that generated summaries capture approximately half of the reference content when accounting for lexical variations. This relatively stronger METEOR performance suggests LLMs produces summaries that maintain semantic fidelity despite using different terminology than human references. The overall temporal decline correlates with increasing document complexity, suggesting potential challenges in maintaining semantic precision as legislative texts grow more elaborate.

5.2. Semantic Level Evaluation

BERTScore (Zhang et al., 2020) was utilized to evaluate the summarization task, providing an assessment of the quality of generated summaries compared to reference summaries. As an automatic metric for text generation, BERTScore calculates a similarity score for each token in the candidate sentence against each token in the reference sentence. This calculation is performed using pre-trained contextual embeddings from BERT models, matching words in both sentences through cosine similarity. The implementation of BERTScore facilitates a comprehensive evaluation of summary quality by capturing semantic relationships beyond lexical overlap, ensuring that generated summaries are accurately compared to their reference counterparts.

BERTScore F1 values above 0.7 generally indicate strong semantic alignment. The range of 0.730-0.757 suggests that LLAMA2-13b produces summaries that capture approximately 73-76% of the semantic content in reference summaries—a reasonably strong performance. These scores significantly outperform the syntactic metrics as ROUGE and BLEU, highlighting BERTScore’s ability to capture semantic rather than lexical similarity, which is particularly important for abstractive summarization. The consistently higher recall versus precision aligns with the observation that LLM summaries are longer than reference summaries. The model effectively captures the reference content but adds supplementary information. The declining trend correlates with increasing document complexity (606 to 1,057 words), suggesting that maintaining semantic precision becomes more challenging as legislative texts grow more elaborate.

Overall, these results indicate good semantic alignment between LLM-generated and reference summaries, especially considering the inherent challenges of legal text summarization. The high recall values demonstrate the model's ability to identify and incorporate key semantic elements, though the precision gap reveals opportunities for generating more focused summaries. The relatively small variation in F1 scores suggests consistent performance despite increasing document complexity over the years.

Tab. 3 – BERTScore semantic evaluation of LLAMA2-13b generated summaries. The table presents yearly semantic similarity metrics between model-generated and human reference summaries. Precision measures how much of the generated content is relevant, Recall quantifies how much of the reference content is captured, and F1-score represents their harmonic mean.

Year	Precision	Recall	F1-score
2023	0.678	0.800	0.730
2022	0.693	0.813	0.747
2021	0.697	0.815	0.750
2020	0.685	0.804	0.739
2019	0.687	0.808	0.742
2018	0.694	0.802	0.743
2017	0.688	0.803	0.740
2016	0.684	0.798	0.735
2015	0.685	0.795	0.735
2014	0.694	0.806	0.745
2013	0.705	0.820	0.757

BERTopic is a topic modeling technique (Grootendorst, 2022) that utilizes pretrained BERT embeddings and class-based TF-IDF (*c-TF-IDF*) to generate interpretable topics and clusters from text. For this experiment, we generated topics for each reference summary (manually created), for summaries produced by the LLM model, and finally, for all topics present in the original text. The objective of this evaluation was to compare the capacity of each summary type to present the same topics identified in the original texts. Through this comparison, we aimed to determine which summary better represents the original text's information: the manual summary or the LLM-generated summary.

The results in Table 4 reveal that BERTopic produces topics that are on par with or superior to human reference summaries in a substantial number of cases, confirming its potential as an unsupervised topic modeling technique that leverages pretrained transformer models. The combined "LLM Wins + Tie" rate consistently exceeds 80% across all years, indicating that LLAMA2 summaries rarely perform substantially worse than human-crafted summaries in capturing topic alignment. For an unsupervised approach competing against human experts in a specialized domain (legislative texts), winning several of direct comparisons while maintaining topic parity in approximately almost of cases represents notable performance. The high tie percentage suggests that LLAMA2 has effectively learned to identify the most significant topics in legislative documents, which is particularly valuable for practical applications in legal information processing.

6. Conclusion

This research investigated the application of Large Language Models (LLMs) for abstractive summarization of Brazilian legislative proposals, examining a comprehensive dataset of over 56,000 texts spanning from 2013 to 2023. The study evaluated the performance of LLAMA2-13b in generating concise and informative summaries of complex legal documents, comparing them against human-created reference summaries using multiple evaluation metrics. The analysis presented provides insights into the capabilities and limitations of LLMs for summarizing legislative documents, particularly Brazilian legislative proposals. The research demonstrates that LLMs can effectively generate coherent and informative summaries of complex legal texts, though with varying degrees of alignment with human-crafted reference summaries depending on the evaluation metric employed.

Our investigation into traditional structural-based metrics such as ROUGE, BLEU, and METEOR revealed sig-

Tab. 4 – BERTopic semantic alignment comparison between reference and LLAMA2-13b generated summaries. The table presents the yearly distribution of topic modeling results using pretrained BERT embeddings and class-based TF-IDF (c-TF-IDF). "Ref. Wins" indicates cases where human-generated reference summaries exhibit stronger topic alignment with original legislative texts. "LLM Wins" represents instances where LLM summaries demonstrate superior topic correspondence. "Tie" shows the percentage of documents where both summary types maintain equivalent topic representation. Note the consistently high tie percentage (65.3%-74.2%) across all years, suggesting comparable semantic performance between human and AI-generated summaries.

Year	Ref. Wins	LLM Wins	Tie
2023	20.8%	17.1%	69.7%
2022	22.6%	13.6%	74.2%
2021	19.5%	15.7%	67.6%
2020	21.5%	18.7%	65.3%
2019	20.3%	16.6%	69.4%
2018	20.7%	16.9%	69.1%
2017	16.1%	13.8%	72.7%
2016	19.2%	15.5%	65.7%
2015	19.3%	18.2%	69.8%
2014	19.3%	15.5%	71.8%
2013	16.1%	13.3%	73.6%

nificant limitations in their applicability to abstractive summarization evaluation. These metrics consistently showed declining performance trends from 2013 to 2023, with ROUGE scores dropping from peaks of 0.399, 0.319, and 0.346 (Unigram, Bigram, and L-score respectively) in 2013 to 0.336, 0.260, and 0.290 in 2023. Similarly, BLEU scores decreased from 0.173 to 0.140, and METEOR from 0.511 to 0.459 over the same period. This decline correlates with the increasing complexity of legislative documents. However, the low performance on these traditional metrics appears to be significantly influenced by the substantial length disparity between reference summaries and LLM-generated summaries rather than actual quality deficiencies. This observation highlights a crucial limitation of lexical overlap-based evaluation for abstractive summarization, where semantic equivalence may exist without exact wording matches.

In contrast, semantic-based evaluation metrics provided more promising results. BERTScore analysis demonstrated strong semantic alignment between LLM-generated and reference summaries. The consistently higher recall versus precision values confirm that while LLM summaries effectively capture reference content, they tend to include additional information. This aligns with our observation regarding summary length differences and suggests that LLMs maintain good semantic fidelity despite using different terminology than human references. Perhaps most interestingly, the BERTopic analysis revealed that LLAMA2-13b generated summaries that were on par with or superior to human reference summaries in a substantial number of cases. The combined "LLM Wins + Tie" rate consistently exceeded 80% across all years, with high tie percentages. This indicates that the LLM can effectively identify and represent the most significant topics in legislative documents, which is particularly valuable for practical applications in legal information processing.

These findings position LLMs as efficient options for summarizing legal documents, with no significant difference between manually and automatically generated summaries in terms of topic representation. However, challenges remain in addressing the complexity of legal terminology and argumentative structures, and in developing evaluation methodologies that adequately assess semantic understanding and preservation of legal validity in generated summaries. Future work should focus on refining prompt engineering techniques to generate more concise summaries that better align with human preferences while maintaining semantic comprehensiveness. Additionally, developing specialized evaluation metrics for legal text summarization that account for domain-specific requirements would enhance the assessment of LLM performance in this context. Finally, exploring fine-tuning approaches specific to legislative documents could potentially improve summarization quality and address the observed performance variations over time.

In conclusion, this study demonstrates the promising potential of LLMs for legislative document summarization while highlighting the importance of semantic-level evaluation metrics in accurately assessing their performance. The results provide a foundation for further research into optimizing LLM applications in legal and governmental domains where efficient information processing is increasingly crucial.

We sincerely appreciate the support provided by the research funding agencies CAPES, FAPESP, and CNPq, whose contributions have been essential to the development of this work. Furthermore, we extend our gratitude to the Brazilian Chamber of Deputies, the Center for Applied Research in Artificial Intelligence Recreating Environments (CPA IARA), and the Brazilian National Institute of Artificial Intelligence (IAIA) for their significant support in advancing this research.

References

Abualigah, L., Bashabsheh, M. Q., Alabool, H., & Shehab, M. (2020). Text summarization: A brief review. *Recent Advances in NLP: the case of Arabic language*, 1–15.

Albuquerque, H. O., Costa, R., Silvestre, G., Souza, E., da Silva, N. F., Vitório, D., Moriyama, G., Martins, L., Soezima, L., Nunes, A., et al. (2022). Ulyssesner-br: A corpus of brazilian legislative documents for named entity recognition. *International Conference on Computational Processing of the Portuguese Language*, 3–14.

Anand, D., & Wagh, R. (2022). Effective deep learning approaches for summarization of legal texts. *Journal of King Saud University-Computer and Information Sciences*, *34*(5), 2141–2150.

Banerjee, S., & Lavie, A. (2005). METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, 65–72. https://www.aclweb.org/anthology/W05-0909

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, *33*, 1877–1901.

Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., & Amodei, D. (2017). Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, *30*.

Egan, N., Vasilyev, O., & Bohannon, J. (2022). Play the shannon game with language models: A human-free approach to summary evaluation. *Proceedings of the AAAI conference on artificial intelligence*, *36*(10), 10599–10607.

El-Kassas, W. S., Salama, C. R., Rafea, A. A., & Mohamed, H. K. (2021). Automatic text summarization: A comprehensive survey. *Expert systems with applications*, *165*, 113679.

Ermakova, L., Cossu, J. V., & Mothe, J. (2019). A survey on evaluation of summarization methods. *Information processing & management*, *56*(5), 1794–1814.

Galgani, F., Compton, P., & Hoffmann, A. (2012). Combining different summarization techniques for legal text. *Proceedings of the workshop on innovative hybrid approaches to the processing of textual data*, 115–123.

Goyal, T., Li, J. J., & Durrett, G. (2022). News summarization and evaluation in the era of gpt-3. *arXiv preprint arXiv:2209.12356*.

Grootendorst, M. (2022). Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Jain, D., Borah, M. D., & Biswas, A. (2021). Summarization of legal documents: Where are we now and the way forward. *Computer Science Review*, *40*, 100388.

Lin, C.-Y. (2004). ROUGE: A package for automatic evaluation of summaries. *Text Summarization Branches Out*, 74–81. https://www.aclweb.org/anthology/W04-1013

Lin, C.-Y., & Och, F. J. (2004). ORANGE: A method for evaluating automatic evaluation metrics for machine translation. *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, 501–507. https://www.aclweb.org/anthology/C04-1072

Liu, Y., Jia, Q., & Zhu, K. (2022). Reference-free summarization evaluation via semantic correlation and compression ratio. *Proceedings of the 2022 conference of the North American Chapter of the Association for Computational Linguistics: human language technologies*, 2109–2115.

Luo, Z., Xie, Q., & Ananiadou, S. (2023). Chatgpt as a factual inconsistency evaluator for abstractive text summarization. *arXiv preprint arXiv:2303.15621*.

Neto, J. L., Freitas, A. A., & Kaestner, C. A. (2002). Automatic text summarization using a machine learning approach. *Advances in Artificial Intelligence: 16th Brazilian Symposium on Artificial Intelligence, SBIA 2002 Porto de Galinhas/Recife, Brazil, November 11–14, 2002 Proceedings 16*, 205–215.

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-j. (2002). Bleu: A method for automatic evaluation of machine translation, 311–318.

Rawte, V., Sheth, A., & Das, A. (2023). A survey of hallucination in large foundation models. *arXiv preprint arXiv:2309.05922*.

Souza, E., Moriyama, G., Vitório, D., de Carvalho, A. C., Félix, N., Albuquerque, H. O., & Oliveira, A. L. (2021). Assessing the impact of stemming algorithms applied to brazilian legislative documents retrieval. *Anais do XIII Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, 227–236.

Souza, E., Vitório, D., Moriyama, G., Santos, L., Martins, L., Souza, M., Fonseca, M., Félix, N., Carvalho, A. C., Albuquerque, H. O., & Oliveira, A. L. (2021, December). An information retrieval pipeline for legislative documents from the brazilian chamber of deputies. DOI: https://doi.org/10.3233/FAIA210326.

Tas, O., & Kiyani, F. (2007). A survey automatic text summarization. *PressAcademia Procedia*, *5*(1), 205–213.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. (2023). Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Vitório, D., Souza, E., Martins, L., da Silva, N. F., de Leon Ferreira de Carvalho, A. C. P., & Oliveira, A. L. (2022). Ulysses-rfsq: A novel method to improve legal information retrieval based on relevance feedback. *Brazilian Conference on Intelligent Systems*, 77–91.

Vitório, D., Souza, E., Martins, L., da Silva, N. F., Oliveira, A. L., de Andrade, F. E., et al. (2023). Building a relevance feedback corpus for legal information retrieval in the real-case scenario of the brazilian chamber of deputies.

von Lucke, J., Fitsilis, F., & Etscheid, J. (2022). Using artificial intelligence for legislation-thinking about and selecting realistic topics. *EGOV-CeDEM-ePart 2022*, 32.

Wang, J., Liang, Y., Meng, F., Shi, H., Li, Z., Xu, J., Qu, J., & Zhou, J. (2023). Is chatgpt a good nlg evaluator? a preliminary study. *arXiv preprint arXiv:2303.04048*.

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2020). Bertscore: Evaluating text generation with bert. *International Conference on Learning Representations*. https://openreview.net/forum?id=SkeHuCVFDr