Intelligent Systems with Applications 31 (2026) 200688

Contents lists available at ScienceDirect

Intelligent Systems with Applications

journal homepage: www.journals.elsevier.com/intelligent-systems-with-applications

Exploring embeddings and document structure recognition for extractive summarization of Portuguese Supreme Court judgments

Martim Zanatti [a,b,*], Rui Melo [c,d], Miguel Won [a,b], Ricardo Ribeiro [a,e], H. Sofia Pinto [a,b], José Borbinha [a,b], João Dias [a,f], Pedro A. Santos [a,b]

[a] INESC-ID, Rua Alves Redol, 9, Lisboa, 1000-029, Portugal
[b] Instituto Superior Técnico, UL, Av. Rovisco Pais, 1049-001, Lisboa, Portugal
[c] Carnegie Mellon University, 5000 Forbes Ave, Pittsburgh, PA, 15213, United States
[d] Faculdade de Engenharia da Universidade do Porto, Rua Dr. Roberto Frias, s/n, Porto, 4200-465, Portugal
[e] Iscte - Instituto Universitário de Lisboa, Avenida das Forças Armadas, Lisboa, 1649-026, Portugal
[f] Faculdade do Algarve, Estr. da Penha, Faro, 8005-139, Portugal

ARTICLE INFO

Keywords:
Embedding models
Extractive summarization
Judgment's segmentation
Legal summarization
Portuguese Supreme judgments

ABSTRACT

Legal text summarization is increasingly relevant due to the growing availability of digital legal documents. In this study, we address the challenges of summarizing judgments from the Portuguese Supreme Court of Justice (STJ) by leveraging domain-adapted embedding models tailored to the Portuguese legal domain, which lacks a robust ecosystem of specialized embeddings. This paper presents the development of domain-specific embeddings for Portuguese legal language, designed to capture the semantics of jurisprudential statements with higher precision. Beyond proposing this resource, the study includes an empirical validation through expert (judge) assessment, demonstrating the embeddings' applicability to real judicial reasoning. Our findings also demonstrate that domain-specific embeddings significantly enhance the quality of extractive summaries, even when employing simple algorithms. Additionally, we propose a segmentation model that identifies the inherent structure and roles of different sections within judgments, a critical step given their extensive length and varying purposes. This model further improves the quality and relevance of generated summaries. Our approach achieved a ROUGE-1 score of **57.23** and a ROUGE-2 score of **25.44**. We evaluated summaries for coherence, completeness, and accuracy using state-of-the-art large language models (Llama and GPT-4o). The LLM-based ratings show limited agreement with ROUGE and with each other, and a cross-evaluation analysis combining these signals with feedback from STJ judges highlights the limitations of any single metric for evaluating legal summaries. Finally, a human evaluation conducted by STJ judges provided valuable insights into the strengths and limitations of our approach.

1. Introduction

Automatic text summarization is the task of generating a summary, a shorter text of a document that expresses the general idea of that document (Zhang et al., 2025). In our work, we focused on the problem referred to as *legal text summarization* (Akter et al., 2025). The increasing availability of digital legal documents has led to a notable rise in research in this area in recent years. Given the typically extensive nature of legal documents, summarization can provide substantial benefits. A judgment is a final decision rendered on a case by a highest court, which can function as a paradigm for resolving analogous cases. In their daily work, lawyers and judges often look up previous judgments to help them understand what is the common practice and understanding a particular problem. Therefore, summaries are particularly important since they are often used to identify and retrieve relevant judgments. The elaboration of summaries by judges varies depending on the judge who writes them. Generally, in specialized domains extractive summarization leads to better results when compared to abstractive summarization (Li et al., 2024).

In parallel, recent research has increasingly focused on leveraging LLMs for legal applications, including recent domain adaptation and benchmarking efforts, some of which target the Portuguese legal system (Canaverde et al., 2025; Colombo, Pires, Boudiaf, Culver et al., 2024; Colombo, Pires, Boudiaf, Melo et al., 2024; Fernandes et al., 2025). These initiatives highlight the growing interest in domain-specific legal AI and the need for dedicated resources. Nevertheless,

* Corresponding author at: Instituto Superior Técnico, UL, Av. Rovisco Pais, 1049-001, Lisboa, Portugal.
E-mail address: martim.zanatti@tecnico.ulisboa.pt (M. Zanatti).

https://doi.org/10.1016/j.iswa.2026.200688
Received 9 April 2026; Received in revised form 25 May 2026; Accepted 11 June 2026
Available online 22 June 2026
2667-3053/© 2026 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC license (http://creativecommons.org/licenses/by-nc/4.0/).

automatic summarization of legal texts in European Portuguese remains an underexplored area, both in terms of datasets and evaluation methodologies.

In this article, we tackle the challenges associated with summarizing judgments from the Portuguese Supreme Court of Justice (STJ). We explore the impact of domain-adapted embedding models specifically fine-tuned for the Portuguese legal domain, which is notably under-resourced in terms of specialized embeddings. By tailoring embeddings to this domain, we demonstrate significant improvements in the quality of extractive summaries, even when leveraging straightforward algorithms. Additionally, we propose a segmentation model that identifies the structure and roles of different sections within judgments. This segmentation allows us to focus exclusively on the most relevant sections, both as context to better understand the judgment and as the basis for extractive summarization. By strategically using specific sections for summary generation, we ensure that the produced summaries are more concise, accurate, and reflective of the judgment's key elements.

Our work makes the following contributions:

- We leverage and extensively evaluate a family of legal embedding models developed for European Portuguese jurisprudence (Melo et al., 2023) comparing multiple fine-tuning variants across a downstream summarization task. To our knowledge, this is the first work to assess the impact of these domain-adapted representations in an end-to-end extractive summarization pipeline, demonstrating significant improvements over the unadapted BERTimbau (Souza et al., 2020) baseline.
- We demonstrate that structural segmentation of legal judgments into their constituent sections is a prerequisite for accurate and concise summarization. By restricting the summarizer to the legally relevant sections — identified automatically by our segmentation model — we achieve substantial further improvements beyond what domain-adapted embeddings alone provide, establishing structural pre-processing as a complementary and independently valuable intervention.
- We propose the first structural segmentation model for Portuguese Supreme Court judgments, defining a taxonomy of document sections in collaboration with STJ judges. Our model reformulates sequence labeling from the word level to the paragraph level, assigning structural roles to each paragraph within a judgment. This design choice reflects the self-contained nature of legal paragraphs and the hierarchical structure of STJ judgments.
- We integrate the segmentation model and domain-adapted embeddings into a unified end-to-end pipeline for extractive summarization of legal judgments, combining structural document understanding with domain-specific semantic representations — two components that, to our knowledge, have not been jointly explored for this language and domain.
- Beyond standard ROUGE evaluation, we assess summary quality using two LLMs (Llama 3.1 and GPT-4o) across three dimensions — Coherence, Completeness, and Accuracy — and further validate selected summaries with (human) judges from the Portuguese Supreme Court of Justice. Building on this multi-faceted evaluation, we present a cross-evaluation analysis that examines where the three signals — lexical, LLM-based, and expert — converge and where they diverge, identifying structural reasons for the observed disagreement and discussing their implications for evaluating legal summaries.

As detailed in the Evaluation section, our best configuration attains competitive ROUGE scores, while LLM-based ratings and feedback from STJ judges highlight the limitations of metrics in evaluating legal summaries. It is important to note that our approach is specifically tailored to European Portuguese legal judgments, and jurisprudence summaries fall outside the scope of this work.

The remainder of the article is organized as follows. Section 2 introduces the context of our problem. Section 3 addresses what has been achieved in this area over the years. Section 4 explains our corpus. Section 5 describes our previous method, which serves as a baseline for assessing the summarization improvements achieved with the Segmenting Judgment model. Section 6 describes our method, the Segmenting Judgment model that identifies the judgments legal sections, the transformation of text in embeddings, and the summarization method used by our system. Section 7 presents and discusses the results of our system, both quantitatively and through an expert validation. Section 8 presents the limitations of the current work. Section 9 concludes and discusses future work.

2. Context of the problem

Our work was developed as part of the *IRIS* project, a collaboration between the INESC-ID and the Portuguese Supreme Court of Justice.[1]

Legal texts have their own characteristics that differ from newspaper articles or scientific texts (Akter et al., 2025). First, legal documents tend to be longer than documents in other domains. In our case this becomes even truer since we are working with judgments of the third instance (Supreme Court) that contains the content of the first instance (District Court) and of the second instance (Court of Appeal). Second, legal documents show a different internal structure when compared with newspaper articles or scientific texts, as explained in Section 4. Third, the vocabulary of legal texts is very specific. Beyond common vocabulary, they have terms that have a specific meaning in the legal field. Finally, citations have an important role in the legal field, as they are used for crediting and referring to authoritative documents and sources. In the case of judgments, there are also guidelines to structure them that may not be followed. In the end, a decision is made by a group of judges, usually three judges, and one of them, usually the rapporteur, produces a summary of that decision. This summary should comply with the following features:

- it has to be smaller than the judgment;
- it has to be informative enough so that judges and lawyers can understand:
  - the content of the judgment;
  - the verdict;
- it is the second intellectual property, the first being the judgment.

Summaries are typically written by the rapporteur, either immediately following the decision or several days later. In the latter case, the rapporteur must reread the judgment, a process that is both time-consuming and labor-intensive. Beyond exploring domain-adapted embeddings and leveraging judgment segmentation to enhance extractive summarization, this study also delivers a practical tool for STJ judges. This tool identifies and returns the most important paragraphs, assisting judges in efficiently producing summaries.

In our case, we focus on extracting the most important paragraphs from the judgments rather than creating a fully finalized summary. We chose this approach due to the following reasons:

- Each judge has a highly specific style when writing summaries;
- There is no standardized summary structure, resulting in summaries that vary significantly in length and writing style;
- Since summaries can assist in similar cases, errors in summary production may lead to incorrect verdicts in analogous legal situations;
- Previous research has demonstrated that extractive summarization yields better results in specialized domains, such as the legal domain (Endres-Niggemeyer & Wansorra, 2004; Ribeiro & Martins de Matos, 2008).

[1] https://juris.stj.pt/pesquisa.

3. Related work

Text summarization, a research area with origins tracing back to the 1950s (Luhn, 1958), has witnessed significant advancements in recent years, largely driven by the advent of transformer architectures (Vaswani et al., 2017) and, more recently, the development of Large Language Models (LLMs) capable of producing high-quality summaries (Basyal & Sanghvi, 2023; Zhang et al., 2024).

A seminal contribution to the use of pretrained embeddings in text summarization was made by Liu and Lapata (2019), who explored the application of pretrained encoders, particularly the BERT architecture (Devlin et al., 2019), to achieve state-of-the-art performance in both extractive and abstractive summarization. For extractive summarization, they introduced a document-level encoder augmented with inter-sentence Transformer layers to identify the most relevant sentences. For abstractive summarization, they proposed an encoder–decoder framework that combined a pretrained BERT encoder with a randomly initialized Transformer decoder. To address the inherent mismatch between the pretrained encoder and the untrained decoder, the authors devised a two-step fine-tuning strategy, which significantly improved performance across multiple benchmark datasets.

Specifically, for the **legal text summarization** problem, Zhong et al. (2019) introduced a novel methodology for summarizing legal decisions. Their approach uses a "train-attribute-mask" pipeline, wherein a Convolutional Neural Network (CNN) model is initially trained on judgments to predict case outcomes. Once trained, the model uses integrated gradients to identify the sentences within the judgment that contribute most significantly to its predictions. These sentences are iteratively masked, and the model is retrained, compelling it to focus on other informative parts of the text. This iterative process ensures the extraction of sentences that are not only predictive of the case outcome but also relevant for summary construction. The final summary is assembled using Maximum Marginal Relevance (MMR) to promote diversification, resulting in a concise yet comprehensive representation of the original judgment.

To summarize German Court Rulings, Glaser et al. (2021) introduced a novel dataset consisting of 100,000 German judgments paired with brief summaries. They pre-trained GloVe (Pennington et al., 2014) to generate word embeddings and employed an attention mechanism to construct sentence embeddings, where the weighted sum of word tokens was directly used as the sentence embedding. For extractive summarization, they trained various models based on Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN), optimizing for the ROUGE (Lin, 2004) score to extract the most relevant sentences from the judgment. For abstractive summarization, they modeled the problem as a conditional language generation task, predicting each word in the summary based on the text of the verdict and previously generated words. They employed an encoder–decoder architecture, where the encoder generated document embeddings from the facts and reasoning sections of the court ruling using attention mechanisms and RNNs, while the decoder combined embeddings from previously generated words and document representations to produce the next word through a softmax layer.

More recently, Huang et al. (2023) proposed a two-stage summarization approach, wherein sentences and keywords are first extracted from legal documents, followed by the generation of an abstractive summary leveraging the extracted content. For extractive summarization, Jain et al. (2024) introduced a methodology that assigns relevance scores to sentences and applies a deep clustering model to refine these scores, thereby enhancing sentence-level relevance. Similarly, Jain et al. (2023) combined statistical and semantic sentence scoring techniques, employing Bayesian optimization to determine optimal feature weights, leading to improved extractive summaries. In the context of abstractive summarization, de Vargas Feijó and Moreira (2023) developed a method tailored to the Brazilian legal domain that mitigates "hallucinations" — the generation of factually inaccurate information by abstractive models — by incorporating textual entailment. This approach involves segmenting the legal document into eight parts, summarizing each segment using a dedicated encoder–decoder model, and selecting the final summary based on the textual entailment score, with the highest-scoring summary designated as the output. Albayati et al. (2025) employed three complementary techniques, a Consecutive-Aware Semantic Voting Mechanism (CASVM), a Weighted Hybrid Sentence Ranking Framework (WHSRF), and a T5-Based Meta-Model (T5-Meta) to enhance the summarization of Turkish court decisions.

Despite significant progress in both extractive and abstractive summarization techniques, challenges remain, particularly in domain-specific applications such as legal text summarization. The integration of advanced language models, tailored attention mechanisms, and optimization strategies has yielded promising results, as evidenced by the variety of approaches discussed. However, issues such as handling domain-specific nuances, mitigating hallucinations, and improving interpretability in abstractive summaries persist. While recent advancements in abstractive summarization have demonstrated significant potential, particularly with methods leveraging encoder–decoder architectures and complex attention mechanisms, this paper opts for a simpler extractive summarization approach using LexRank (Erkan & Radev, 2004). This choice is motivated by several factors: first, it allows us to better explore the role of domain-adapted embeddings for the Portuguese legal domain; second, it facilitates the investigation of whether segmentation of judgments positively influences summary generation; and finally, given that this work will also result in a tool intended for judicial use, we prioritize methods that avoid the risk of generating hallucinations, which are a known challenge in abstractive summarization techniques.

LexRank is an algorithm for extractive text summarization that leverages the concept of centrality within a graph structure. The method evaluates the centrality of each sentence in a cluster, selecting the most important ones for the summary. Centrality, in this context, is derived from the concept of prestige in social networks, where a sentence that exhibits high similarity to numerous other sentences in the document set is considered more central to the topic. LexRank calculates the similarity between sentences using the IDF-modified cosine similarity, resulting in a similarity matrix in which each entry represents the similarity between a pair of sentences. Consequently, sentences from documents that deviate from the main topic are less likely to receive high scores, as they are not supported by local votes. An enhancement to the original LexRank algorithm, known as Continuous LexRank, incorporates the strength of similarity links by using cosine similarity values directly to construct the graph. In this variation, the LexRank scores of linking sentences are weighted by the corresponding similarity values, further refining the centrality calculations.

4. Dataset

Our corpus contains 700 judgments of the Portuguese Supreme Court of Justice and their respective summaries. While the original judgments include confidential information and are not publicly available, anonymized versions exist.[2] This does not impact the reproducibility of the present work, as it concerns jurisprudential statements in which specific entities are never mentioned. The smallest judgment of our corpus has only 248 words while the largest one has 66,988 words. The average size of judgments is 7760 words. Moreover, 61% of the judgments have less than 7760 words, but within the judgments with less than 7760 words, 58% of them have more than 3880 words.

[2] https://juris.stj.pt/pesquisa.

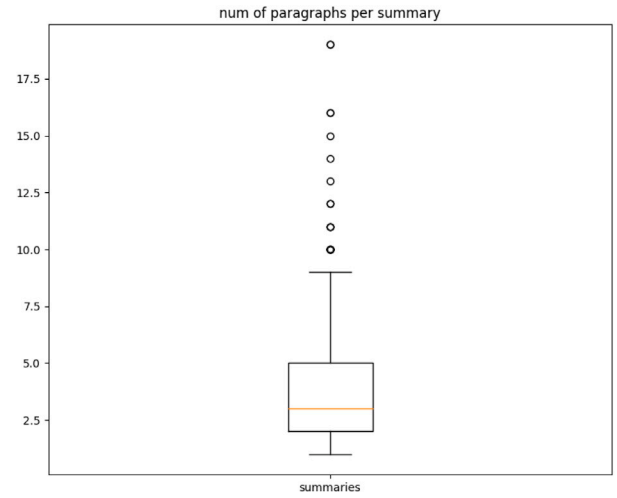


Fig. 1. Boxplot showing the distribution of paragraphs per summary. The minimum is 1, and the median is ≈3. The maximum observed is ≈20, though the upper quartile (Q3) ends at ≈9, indicating that 8 summaries with more than 10 paragraphs are considered outliers. Outliers are defined as values above Q3 + 1.5 × IQR.

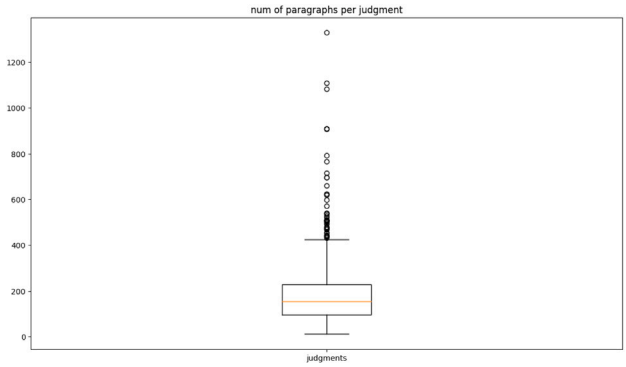


Fig. 2. Boxplot showing the distribution of the number of paragraphs per judgment. The minimum value is below 100, with a median of ≈150. The interquartile range extends up to ≈250, and outliers appear beyond ≈400 paragraphs, with a few extreme cases exceeding 1000. Outliers are defined as values above Q3 + 1.5 × IQR.

Regarding the summaries, the smallest one has only 17 words (see Table 1, count before translation), while the largest one has 1523, which highlights the differences in size and writing style between summaries. The average summary has 220 words, where 61% of the summaries have fewer words than the average. However, 90% of the judgments that have more words than the average have a corresponding summary with more words than the average. Given the difference in size between summaries as can be seen by the difference in the number of words and the difference in the number of paragraphs (see Fig. 1), it does not make sense to suggest the same number of paragraphs for every judgment.

With this in mind, we observe that choosing the number of paragraphs to be extracted depends on the length of the judgment. Therefore, we looked into the judgments in the 1st and 3rd quartiles of Fig. 2, to avoid high and low values, and we calculate the average of the ratio between the paragraphs of the summaries and the paragraphs of their respective judgments, which is 2%. Therefore, the number of paragraphs extracted from the judgment to produce the summary corresponds to 2% of the total number of paragraphs in the judgment.

Concerning the structure of judgments, an extensive analysis of the judgments and meetings with the Portuguese Supreme Court judges showed that a judgment is divided into three main sections. The first section corresponds to the report (*Relatório* in Portuguese) section, (1) where the content of the judgment is introduced, (2) what the various invested parties are seeking, and (3) which decision is to be considered and decided. The second section corresponds to the grounds (*Fundamentação* in Portuguese) section. Usually, this section is divided into two sub-sections: facts (*Matéria de Facto* in Portuguese) and law (*Matéria de Direito* in Portuguese). The first sub-section sets out the proven facts of the case while the second sub-section argues about everything that has been elaborated so far to make the final decision. The last section corresponds to the decision section. Typically, this section is small, as it only contains the final decision.

Regarding summaries, there is no standard structure. There are many examples where the summary consists of only one or two sentences (see Table 1), and there are many examples as in Table 2, where the summary is larger, structured, and listed with numbers. We find that judges typically produce the summaries in an extractive way, i.e., they extract sentences or paragraphs directly from the judgments. After this extraction they usually introduce minor changes, mostly to rephrase the extractive sentences, but in some cases there are no changes at all as can be observed in Table 3, which shows sentences or paragraphs taken from judgments that are equal to those in the respective summary. To conduct a more systematic study on the extractiveness of the summaries relative to the judgments, we employ the ROUGE metric (Lin, 2004), to measure the number of overlapping *n*-grams between the summary and the respective judgment. As shown in Table 4, the ROUGE-1 and the ROUGE-L achieved almost the maximum score. The ROUGE-2 results are also notably high. These high

Table 1
Summary example 1.

Summary example 1
An appeal to the Supreme Court is not admissible when its sole basis is the allegation of procedural nullities in the judgment under appeal.

Table 2
Summary example 2.

Summary example 2
I. Within the scope of a medical service contract, based on a surgical extraction procedure, the medical professional assumes an obligation of result regarding the referred extraction with local anaesthesia, and an obligation of means regarding the application of the appropriate and adequate technique for that result, as well as with respect to the performance involving the same technique, in accordance with the rules of medicine accepted and followed by the specialized community (leges artis) of medical practice and the combination of those rules with the specific scientific knowledge required of the physician and his or her experience. II. The typical obligations of the medical contractual party are governed by the general principle of contractual liability, as provided in art. 798(1) of the Civil Code, as well as by the regime of judicial procedure in art. 799(1) of the Civil Code. A distinction is drawn between the absolute subjective right of the counterparty (art. 70(1) CCiv.; arts. 25(1) and 29(1) CRP) or a legal norm protecting third-party interests, with the CRP framed within the so-called "new regime" of civil contractual/non-contractual liability — breach or defective performance — and extra-contractual civil liability (covered by art. 483(1) of the Civil Code). In cases of breach of a medical service contract, the latter liability must always be considered, since contractual liability lies elsewhere (and, in this sense, there have been concrete cases of responsibilities of different nature), along with the consequent compensation for non-pecuniary damages, without prejudice to being weighed, depending (in a hybrid method of combination) on the rules of delimited liability within what is seen in the extension of servitude and in the reparability of the civil injuries/sequelae caused, where its particular unlawfulness does not "negate" the present motivational framework of art. 488 of the Civil Code. III. Contractual liability applies when there is an "obligation of means" (and not of result and means) requiring the adoption of all adequate/appropriate procedures and means for performing the service (arts. 762(1) and 763(1) CCiv) or in cases of defective performance and/or insufficiency for the intervention attributable to the physician according to normative terms and techniques used (by reason of conformity with the rules of leges artis) to obtain the agreed result for the treatment/intervention.

Table 3
Comparison between paragraphs in the summary and paragraphs extracted from the judgment (translated for English for readability purposes).

Paragraphs		
Summary paragraph	Judgment paragraph	Section
However, not every form of review conducted in annulment proceedings can lead to the annihilation of an arbitral award merely because the judge deems the reasoning to be insufficient or unconvincing.	However, it underscores that not every form of review conducted in annulment proceedings can lead to the annihilation of an arbitral award merely because the judge deems the reasoning to be insufficient or unconvincing.	Grounds - law
What becomes apparent is the Defendant's violation of a credit right held by the Plaintiff against the Bank, allowing the claim to be characterized as an action aimed at enforcing liability based on tort, falling instead within the scope of bullet 2 of the aforementioned Article 71.	Indeed, what becomes apparent here is the Defendant's violation of the Plaintiff's credit right against the Bank, which allows for the characterization of the claim as an action aimed at enforcing liability based on tort, rather falling within paragraph 2 of the aforementioned Article 71.	Grounds - law
The Plaintiff's statement, as found in the final part of the initial petition, indicating that the action was brought under Article 279, bullet 2, of the CPC (in this case, one day before the end of the 30-day deadline following the final judgment that acquitted the Defendant in the instance), does not provide grounds for the Defendant's urgent summons, as per Article 561 of the aforementioned code.	The 30-day period stipulated in bullet 2 of Article 279 of the CPC for the commencement of the new action or the Defendant's summons, with the possibility of using the evidence presented in previous proceedings. It was further contested that there was no request in the Plaintiff's petition for the Defendant's urgent summons since there was no justification for it, as provided under Article 561 of the CPC.	Report

Table 4
ROUGE scores of the summaries and their respective judgments. We consider 1-gram (ROUGE-1), 2-grams (ROUGE-2), and the longest common sequence (ROUGE-L).

ROUGE-1	Median	ROUGE-2	Median	ROUGE-L	Median
0.96	0.98	0.81	0.84	0.95	0.96

ROUGE scores indicate that nearly every word or sequence of words in the summaries can be found in their corresponding judgments. We speculate that the scores of ROUGE-1 and ROUGE-2 are not higher due to the incorporation of connector words, which are used to enhance the readability of the summaries. The score of ROUGE-L indicates that paragraphs of the judgments are used to write the correspondent summaries. Given these results, it is clear that using extractive summarization is an effective approach for producing valuable and relevant summaries for Portuguese Supreme Court judgments. Furthermore, considering that extractive summarization avoids generating "hallucinations", this method proves particularly advantageous in the legal domain.

5. Pre-processing

Before implementing and training the Segmenting Judgment model, we employed a more basic method to pre-process judgments for extractive summary generation. We describe this method, which serves as a baseline for evaluating how the Segmenting Judgment model enhanced extractive summarization (see Section 7). The rule-based pre-processing pipeline operates independently from the Segmenting Judgment model introduced in Section 6.1. While pre-processing serves as a heuristic baseline, the segmentation model relies on its own supervised annotation scheme and is trained separately.

This study employs a BiLSTM-CRF model to segment the judgments into predefined structural sections. Although Portuguese Supreme Court judgments formally follow a prescribed structure (typically comprising sections such as Report, Grounds, and Decision), in practice the textual realization of these sections is far from uniform. Variations in drafting style, formatting, and even section titling make automatic segmentation a non-trivial task. Simple rule-based heuristics are insufficient to capture this variability, especially when section boundaries are implied rather than explicitly marked. For this reason, we developed a data-driven BiLSTM-CRF model capable of learning implicit structural cues beyond explicit headers, enabling a more robust and generalizable segmentation of judgment texts.

As delineated in Section 4, it is customary for judges to extract specific segments from judgments for summary creation. Predominantly, the "law" sub-section from the "grounds" section is extracted, a pattern clearly illustrated in Table 3. With this insight, we subjected the judgments to pre-processing to facilitate the summarizer in extracting pertinent paragraphs from crucial sections. All methodologies detailed in the Pre-processing section pertain to our previous research endeavors.

```
<p>Código de Processo Civil Anotado,VI,336.<a class="footnote-back" href="#fnref1"></a></p>
```

Fig. 3. Footnote HTML example.

We divide each judgment into paragraphs. The pre-processing phase involves distinguishing between paragraphs that can be included in the summary and those that should be excluded. Table 5 outlines the types of paragraphs that are considered summarizable and those that are not. In the following sections, we provide a detailed description of the pre-processing steps.

5.1. Section titles

To check if a paragraph is a section title we take advantage of the fact that the section titles are uniform. Therefore, we create a list of the section titles we found, and if a paragraph is less than five words, does not end with a punctuation mark, and the text is a section title contained in the list, the paragraph is considered a section title. We are aware that if there are section titles that are not contained in the list, the system is not able to identify them. The enhancements have been elucidated by incorporating the section title in the Segmenting Judgment model, as detailed in Section 6.1.

5.2. Footnotes

HTML format text has information about footnotes as can be seen in Fig. 3. If a paragraph has the footnote class, the paragraph is considered not summarizable.

5.3. Italic text

Citations (that are in *italic*) are used for crediting and referring to authoritative documents and sources as stated in Section 2. This type of text is not used to write a summary. However, there are paragraphs that contain some words in *italic* that are used in summaries as we observed in the ground truth summaries. To differentiate between these two types of paragraphs, we consider a paragraph to be a citation — and thus not summarizable — if and only if 90% of its words are in *italic*. We also check the paragraph preceding a citation paragraph. If it ends with the punctuation mark colon (":") the paragraph is considered not summarizable.

Table 5
Characteristics of paragraphs that are considered summarizable or unsummarizable.

Paragraph type	Summarizable
Title section paragraph	False
Italic paragraph	False
Paragraphs containing only symbols	False
Small paragraphs	False
Paragraphs that end with the ":" punctuation mark and preceding an italic paragraph	False
Paragraphs after final date paragraph	False
Paragraphs that do not contain any of the above specifications	True

5.4. Symbols and small paragraphs

Paragraphs that only contains symbols or it has less than four words is considered not summarizable.

5.5. Final date

After the decision section, the date of the judgment is written, followed by the names of the judges. In some judgments, after the names of the judges comes the explanation of vote. The explanation of vote is an instrument that allows the judges to explain the reasons for their vote on a given matter, after the result of the vote has been proclaimed. The final date and all that is written after it is not relevant to the summaries. To find the final date we trained a Named-Entity Recognition model. Our NER model was trained with spacy-transformers,[3] where we use the Legal-BERTimbau (see Section 6.2.2) model to create the embeddings. We used two datasets to train the model:

- CINTIL dataset that has the PER (person), LOC (location) and ORG (organization) entities annotated[4];
- Sigarra dataset that has the PER, LOC, ORG and DAT (date) entities annotated.[5]

The final date can be expressed in three different ways: location (LOC) followed by the date (DAT), organization (ORG) followed by the date (DAT), or only date (DAT). If the final date is found, this paragraph and all that comes after are considered not summarizable.

6. Architecture

The architecture of our system is illustrated in Fig. 4. The system takes as input a judgment from the Portuguese Supreme Court. All judgments in our corpus are Microsoft Word[6] documents in DOCX format. For each judgment, we segment the text into paragraphs using the Python package beautifulsoup4.[7] The Segmenting Judgment model (described in Section 6.1) then assigns each paragraph to its corresponding section.

6.1. Segmenting judgment model

Following Huang et al. (2015), who were pioneers in applying the Bidirectional LSTM-CRF model to NLP sequence tagging tasks, we adopt a similar architecture for our segmentation model. The Bidirectional LSTM captures sequential context from both forward and backward directions, while the subsequent CRF layer leverages these representations to produce globally normalized tag sequences, ensuring the coherence of the predicted labels.

Our adaptation differs from the original in two key respects. First, rather than tagging individual words in sentences, we operate at the paragraph level, where each paragraph is assigned a section label. While words in a sentence are closely interrelated, paragraphs in legal judgments are more self-contained units that nonetheless follow a predictable hierarchical structure — a property that makes this adaptation effective. This structural awareness is central to our summarization approach: by identifying the most relevant sections of each judgment, we can discard paragraphs that are not pertinent to summarization, enabling a more targeted and efficient extraction of key content. Second, padding is applied at the document level to ensure uniform paragraph counts across judgments, and the tag set is replaced with section labels specific to Portuguese STJ judgments, rather than generic named entity tags. Despite these differences, the underlying principle remains consistent: assigning specific tags to sequential units to capture the essential structure of the text.

It should be noted that the segmentation model employed in this study corresponds to an earlier version of our system. A subsequent iteration, described in Zanatti et al. (2024), introduces additional refinements and improved performance. Nevertheless, given the dependency of the present analysis on judge-based evaluations and experimental consistency, we report here the results obtained with the original version of the model.

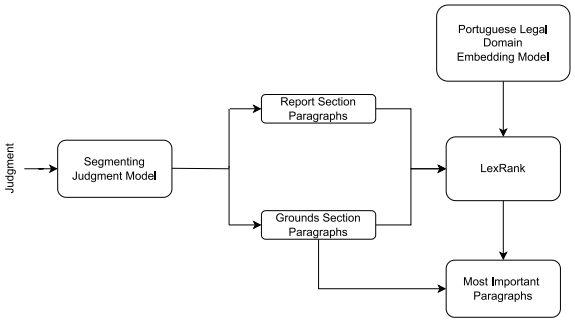


Fig. 4. System Architecture. The judgment is divided into several sections (see Section 6.1). The Report and Grounds sections are processed by LexRank, which identifies and returns the most relevant paragraphs from the Grounds section to generate the extractive summary.

6.1.1. Sections

In this section, we delineate and elaborate on the tags or sections employed in our architecture. Apart from the three primary sections that consistently appear in the judgments — namely the Report (referred to as *relatório* in Portuguese), Grounds (*fundamentação*), and the Decision section (detailed in Section 4) — we incorporate six additional sections. These are:

- Header (*Cabeçalho*);
- Delimitation (*Delimitação*);
- Signature (*Colectivo*);
- Declaration (*Declaração*);
- Footnotes;
- Title.

[3] Version of spacy=3.2.3.
[4] http://cintil.ul.pt/pt/cintilfeatures.html#acquire.
[5] https://rdm.inesctec.pt/dataset/cs-2017-004.
[6] https://www.microsoft.com/en-us/microsoft-365/.
[7] Version of beautifulsoup4=4.11.1.

The **Header** section serves as the introductory section of the judgment. It encompasses the following key elements:

- **Document Identification**: This provides information that designates the document as a legal judgment from a particular court or tribunal.
- **Case Reference Number**: A distinct identifier for the judgment or case, this number is pivotal for record-keeping and future referencing.
- **Type of Procedure**: This indicates the nature of the case or legal procedure, encompassing categories like criminal, civil, administrative, among others.

While the **Delimitation** section does not appear in most judgments of the Portuguese Supreme Court, it plays a pivotal role in setting the stage for the judgment's primary content. This section delineates the specific matters the court or tribunal intends to address, defining the scope of the case, elucidating the primary legal questions, and pinpointing any specific issues or concerns the court will evaluate.

The **Signature** section typically lists the names of the three judges who contributed to the case. Additionally, it includes the date, as detailed in Section 5.5.

The **Declaration** section serves to offer official statements or clarifications from the judges regarding their decision. This section might include:

- **Judgment Affirmation**: An official statement affirming the court's decision regarding the matter at hand.
- **Dissenting Opinions**: If any of the judges disagree with the majority opinion, they might include a declaration explaining their reasons for dissent.
- **Clarifications**: Judges might use this section to provide clarity or additional insight regarding specific aspects of their decision or reasoning.

Footnotes in legal documents, especially in judgments, carry substantial weight and diverge from the traditional usage of footnotes found in other genres of documents. Their inclusion fulfills several essential roles:

- **Legal Citations**: One of the primary uses of footnotes in legal documents is to cite relevant legal authorities, such as statutes, regulations, case law, and other legal sources. These citations provide the foundation upon which legal arguments and conclusions are based.
- **Clarification**: Sometimes, a statement in the main text might need additional context or clarification without disrupting the document's flow. Footnotes offer a space to elaborate on terms, concepts, or provide additional insights without interrupting the primary narrative.
- **Historical Context**: Footnotes can provide historical background on specific legal principles, past decisions, or statutes.
- **Additional References**: Besides legal citations, footnotes might also direct readers to secondary sources, such as legal commentaries, journals, or scholarly articles, providing a broader understanding or different perspectives on the matter at hand.
- **Cross-referencing**: Especially in lengthy judgments, footnotes can be used to cross-reference other parts of the document, making it easier for readers to navigate complex arguments or multifaceted discussions.

Within the structural framework of Portuguese judgments, the initial segment is recognized as the **Header** section. It is worth noting that this section is not a universal feature across all judgments issued by the Portuguese Supreme Court. Following this, the **Report** section consistently presents itself in every judgment. In certain judgments, a **Delimitation** section can be found nestled between the Report and the **Grounds** sections, the latter of which is an omnipresent feature in the judgments. The **Decision** section succeeds the Grounds section and is another constant element. Subsequent sections include the **Signature** and **Declaration** sections; however, their appearance is not consistent across all judgments and their sequence can vary. Additionally, while some judgments designate specific **titles** to particular sections, the presence of **footnotes** is also observed in select judgments.

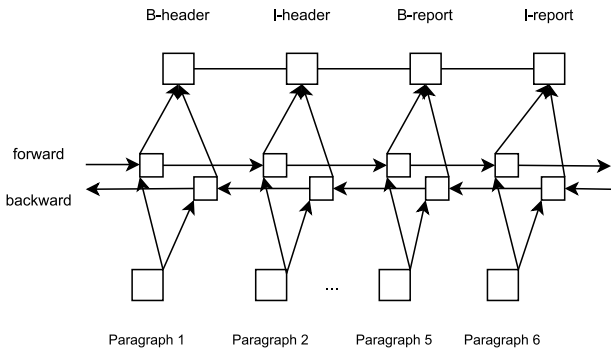


Fig. 5. Segmenting Judgment Architecture. Here, the transition from the Header section to the Report section is illustrated.

In Fig. 5, the architectural blueprint of the Segmenting Judgment Model is depicted. Our method aligns with the approach by Zhiheng Huang et al. In this structure, each section (excluding the title tag) bifurcates into a B-section and an I-section. Herein, the B-section signifies the commencement of the respective section, while the I-section denotes its intermediate position.

6.1.2. Dataset

To effectively train our Segmenting Judgment Model, a dataset segmented into the specified sections was essential. However, no such dataset existed for the Portuguese Supreme Court's judgments, necessitating its creation from scratch. This process was conducted semi-automatically: sections with explicit titles facilitated automated segmentation, which was then validated by a script to ensure accurate text-section associations. Any detected discrepancies were manually corrected. The resulting dataset comprises 670 distinct judgments. While our architecture was trained on these judgments, we observed that they largely followed a uniform structure. As a result, the model struggles to generalize effectively when encountering judgments that deviate from this established pattern.

To address this issue, we created variations of the original 670 judgments by omitting specific sections that were not ubiquitous in all judgments, namely: Header, Delimitation, Signature, Declaration, Footnotes, and titles. This yielded a diversified dataset of 5340 judgments, encompassing various structural variations. With this enriched dataset, the model is now better poised to process judgments of diverse structural types.

6.1.3. Training procedure

We hold out 20% of the original judgments as an untouched test set and use the remaining 80% for model development. Since our dataset contains multiple structural variants derived from each original judgment, we perform group-wise splitting by original judgment ID for both the held-out test set and cross-validation, ensuring that no variants from the same judgment appear across splits or folds. Hyperparameter optimization is conducted on the development portion using 10-fold *GroupKFold* cross-validation, which is deterministic and requires no random seed. For the final model, the development set is further partitioned into training and validation subsets (90%/10%) using *GroupShuffleSplit* with *random_state=42*. The best-performing configuration trains for up to 10 epochs with a batch size of 64, using the

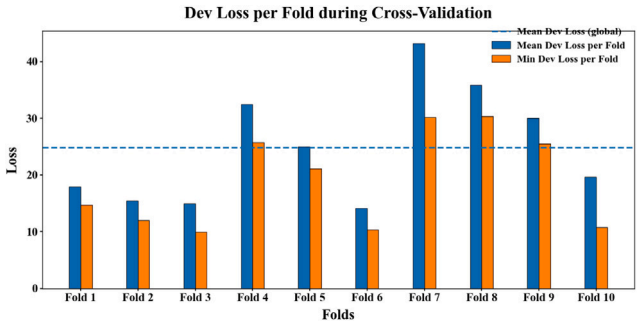


Fig. 6. Validation loss per fold, showing the mean loss for each fold, the overall average across folds, and the minimum validation loss recorded in each fold.

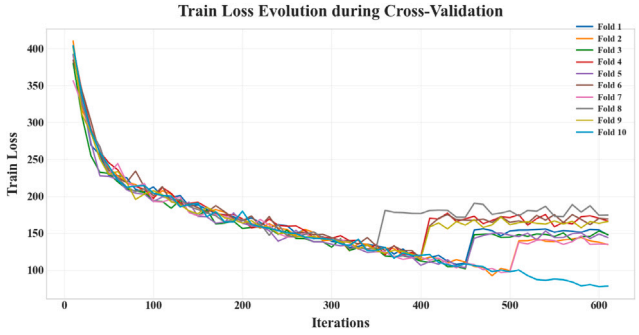


Fig. 7. Training loss evolution across 10 epochs for each of the 10 cross-validation folds.

Adam optimizer with a learning rate of 10^{-3}, one hidden layer with 1024 units, a dropout rate of 0.5, and early stopping with patience 4.

We employ the *stjiris/bert-large-portuguese-cased-legal-mlm-nli-sts-v1* model[8] to encode paragraphs into embeddings. This model, a legal adaptation of BERTimbau large, represents each paragraph in a 1024-dimensional dense vector space. Further details on the embedding model variants are provided in Section 6.2.2.

Figs. 6 and 7 show the validation and training loss curves across folds. We observe noticeable variability in validation loss across folds, with the minimum per-fold loss consistently below the mean, indicating that the model reaches better performance at specific training checkpoints. Training loss decreases sharply in early epochs across most folds, with occasional plateaus in some splits, likely reflecting variation in judgment structure and label distribution. We use cross-validation primarily as a diagnostic for training robustness; for the final system, we train on the full development set and select the checkpoint via validation loss.

On the held-out test set, the model achieves an **accuracy** of **0.9417**, a **precision** of **0.9138**, a **recall** of **0.8033**, and an **F1-score** of **0.8347**, supporting the robustness of the proposed segmentation model.

Among the 205,352 paragraphs in the test dataset, the model misclassified 11,973 paragraphs. Fig. 8 illustrates the distribution and nature of these errors. Because every section — except for title — is represented by a "B-" and "I-" label, two distinct types of errors emerge.

The first type corresponds to intra-section errors, where the model correctly identifies the section but confuses the paragraph's position within it (e.g., predicting I-grounds instead of B-grounds). These errors reflect boundary ambiguities rather than structural misunderstanding. The second type consists of inter-section errors, where a paragraph is assigned to an entirely incorrect section. When excluding intra-section errors, the model misclassified 11,636 paragraphs.

As shown in Fig. 8, the most frequent inter-section confusion involves Footnotes and Grounds. This is plausible in the legal domain: footnotes often contain substantive references, quotations, or short argumentative remarks that can resemble the legal reasoning typically found in the Grounds section, making them harder to separate based purely on local textual cues. In addition, we observe a non-negligible amount of confusion between Report and Grounds, particularly around their transition, where the model may delay the onset of Grounds or prematurely assign reasoning-like paragraphs to Report. Excluding Footnotes vs. Grounds, the remaining inter-section confusions occur predominantly between adjacent sections, suggesting that the model captures the overall document structure and that most errors concentrate around local boundary decisions rather than arbitrary section swaps.

To further characterize the practical impact of these errors, we analyzed them at the document level across the 1064 judgments in the test set. Overall, 459 judgments (43.1%) contain at least one misclassified paragraph. However, this aggregate figure conflates errors of different severity. Decomposing by error type reveals a more nuanced picture: intra-section errors affect 295 judgments (27.7%), involving 337 paragraphs where the section identity is correct but the boundary marker (B- vs. I-) is mispredicted. Inter-section errors are more prevalent, affecting 409 judgments (38.4%) and accounting for 11,636 misclassified paragraphs.

Since the summarization pipeline prioritizes the Grounds section — as emphasized by STJ judges and further supported by our evaluation results (Section 7), where the Report section was shown to contribute meaningful contextual information — not all inter-section errors carry equal weight for downstream summarization. We therefore distinguish between two types of summarization-relevant errors: *lost paragraphs*, where the ground-truth label is Report or Grounds but the model predicts a different section, causing relevant content to be excluded; and *spurious paragraphs*, where the ground-truth label belongs to another section but the model predicts Report or Grounds, introducing extraneous content into the summarization input. Lost paragraphs affect 231 judgments (21.7%), comprising 3714 paragraphs that the summarizer never receives. Spurious paragraphs are more frequent, affecting 274 judgments (25.8%) and introducing 10,952 extraneous paragraphs.

Examining the confusion matrix (Fig. 8) reveals the origin of the majority of spurious paragraphs. The largest source corresponds to Foot-note paragraphs misclassified as I-Grounds (6556 paragraphs). As discussed in Section 6.1.1, footnotes in the legal domain frequently contain substantive references, quotations, and argumentative remarks that closely resemble the reasoning found in the Grounds section. Combined with the fact that Grounds is by far the most content-rich section in STJ judgments, the model appears to resolve this ambiguity in favor of the dominant label, systematically absorbing footnote content into Grounds. A secondary but non-negligible source involves Report paragraphs misclassified as I-Grounds (2031 paragraphs). This error is arguably less severe: since Report content is itself intended for summarization, these paragraphs would not introduce entirely off-topic material. However, they may subtly distort the summarization input by displacing genuinely argumentative content, potentially leading the extractive model to select paragraphs that reflect factual reporting rather than legal reasoning, with a negative effect on summary quality.

To better isolate the true error rate from the inflation introduced by the eight structural variants per judgment, we restrict the analysis to the 134 original judgments in the test set. Of these, 32 (23.9%) contain at least one misclassified paragraph, a notably lower rate than the 43.1% observed across all variants. This reduction is expected: the structural variants, by removing sections that serve as contextual anchors for the overall document layout, increase the difficulty of correct boundary classification in the remaining content. Nevertheless, these variants play a crucial role in evaluating the robustness of the

[8] https://huggingface.co/stjiris/bert-large-portuguese-cased-legal-mlm-nli-sts-v1.

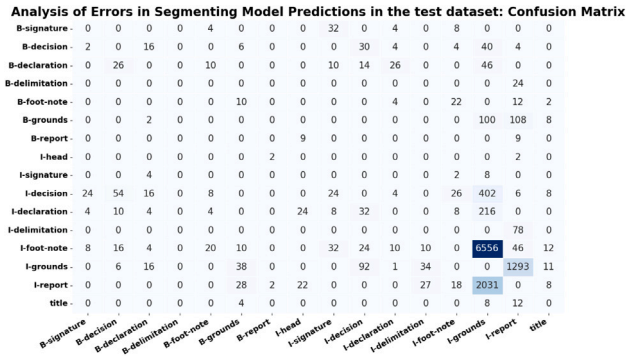


Fig. 8. Confusion matrix providing an in-depth analysis of errors made by the Segmenting model in predicting judgment sections. In each cell is the frequency of misclassifications, showing only the errors made by the model. For instance, the cell at row "B-grounds" and column "I-report" indicate the number of times the model incorrectly predicted the "I-report" instead of "B-grounds".

segmentation model, ensuring that it performs reliably across judgments with different structural configurations rather than relying on the presence of all sections to produce correct predictions.

Decomposing by error type, intra-section errors affect 20 judgments (14.9%), involving only 23 paragraphs — a negligible impact on summarization quality. Inter-section errors are more consequential, affecting 27 judgments (20.1%) and misclassifying 1252 paragraphs. Of the 32 judgments with at least one error, 14 exhibit only intra-section or inter-section errors between sections irrelevant to summarization, and are therefore unlikely to affect summary quality. The remaining 18 judgments (13.4% of all original judgments) contain errors with direct summarization impact: 9 judgments lose relevant content (281 paragraphs where Report or Grounds are misclassified as another section), while 15 judgments receive spurious content (1196 extraneous paragraphs incorrectly assigned to Report or Grounds). The pronounced asymmetry between spurious and lost paragraphs — more than four to one — reinforces the pattern observed in the full dataset: the model systematically over-predicts Report and Grounds labels, likely driven by their dominance in the training data. In Section 7.1.1, we directly assess whether this asymmetry translates into measurable degradation of summarization quality for the affected judgments.

6.2. Word and sentence embeddings

Word embeddings implement the Distributional Hypothesis (DH), which posits that words appearing in similar contexts share similar meanings. Each word is represented as a vector in a multidimensional space, and semantic similarity between words can be measured via cosine similarity between their respective vectors. Since these representations form a linear space, arithmetic operations in the embedding space are also meaningful.

Several implementations exist, including Word2Vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014), both of which use neural networks for unsupervised training but differ in their underlying formulation.

6.2.1. Word2Vec

Word2Vec (Mikolov et al., 2013) is a two-layer neural network based on the idea that words appearing in similar contexts share similar meanings. It supports two training objectives — CBOW and Skip-Gram — of which we adopt the latter, which predicts context words from a target word. We trained our Skip-Gram model using the gensim package[9] on a corpus of 61,239 judgments and summaries from *DGSI*,[10] using lemmatized tokens with stopwords removed,[11] producing 100-dimensional word vectors.

6.2.2. Legal-BERTimbau

Legal-BERTimbau (Melo et al., 2023) is a fine-tuned BERT model designed for Semantic Textual Similarity (STS). Bidirectional Encoder Representations from Transformers (BERT), introduced by Google AI Language (Devlin et al., 2019), leverages the groundbreaking Transformers architecture. This paper explores multiple variants of Legal-BERTimbau. The model is derived from BERTimbau (Souza et al., 2020). While BERTimbau is already a pretrained language model tailored to the Portuguese language, further adaptation was required for the legal domain. This fine-tuning stage was crucial to ensure that the model generating the embeddings could effectively interpret and represent judicial texts.

Data Legal documents contain specialized language that is not commonly found in conventional websites or books. Members of Project *IRIS* conducted the data collection process using *ecli-indexer*.[12]

Ecli-indexer is a repository containing multiple tools for extracting documents from *DGSI*, a public legal database, and indexing them into *ElasticSearch*. The retrieval process gathered HTML content from multiple web domain pages containing legal judgments, resulting in a total of 31,690 judgments.

To train, test, and validate the models, the dataset was divided into three subsets with the following proportions: 80% for training, 10% for testing, and 10% for validation. This dataset was published on the Hugging Face platform[13] to enhance model reproducibility and support future research efforts. The specific dataset divisions are as follows:

- Training dataset — 26,952 judgments;
- Testing dataset — 3169 judgments;
- Validation dataset — 3169 judgments.

Domain adaptation BERTimbau was originally trained on BrWaC, a Brazilian Portuguese web corpus that differs significantly from European Portuguese legal text in both register and vocabulary. Domain adaptation was therefore necessary to align the model with the specialized jargon and structured language of Portuguese jurisprudence. The adaptation experiments were conducted on a NVIDIA GeForce RTX 3090 24 GB GPU.

Domain adaptation was performed using two approaches: BERT's Masked Language Modeling (MLM) and the Transformer-based Sequential Denoising Auto-Encoder (TSDAE). MLM exposes the model to domain-specific vocabulary and jargon, while TSDAE trains sentence embeddings by encoding corrupted sentences into fixed-size vectors and requiring the decoder to reconstruct the originals, encouraging semantically rich representations. A significantly lower learning rate was used compared to pre-training to prevent overfitting, given the model's parameter count and the relatively small dataset size.

- stjiris/bert-large-portuguese-cased-legal-mlm[14];
- stjiris/bert-large-portuguese-cased-legal-tsdae.[14]

Semantic textual similarity Semantic Textual Similarity (STS) is a regression task that quantifies the similarity between two text segments on a numeric scale from 1 to 5. We developed Sentence-BERT (SBERT) versions of each Legal-BERTimbau variant by adding an independent linear layer and fine-tuning with a mean squared error loss. Training

[9] Version of gensim=4.2.0.

[10] https://www.dgsi.pt/.

[11] The words classified as stopwords are those defined by the NLTK (Bird et al., 2009) stopwords package for the Portuguese language, nltk=3.7.

[12] https://github.com/diogoalmiro/ecli-indexer.

[13] https://huggingface.co/datasets/stjiris/portuguese-legal-sentences-v0.

[14] https://huggingface.co/stjiris.

used three Portuguese STS datasets: ASSIN[15] (Fonseca et al., 2016) (5000 training pairs), ASSIN2[16] (Real et al., 2020) (6500 training pairs), and the Portuguese subset of stsb_multi_mt (May, 2021) (5749 training pairs), yielding a total of 20,197 sentence pairs for fine-tuning.

Natural language inference According to Reimers and Gurevych (2019), SBERT, a slight improvement in STS performance was observed when models were trained with Natural Language Inference (NLI) data. The previously utilized ASSIN and ASSIN2 datasets also include information regarding the degree of relatedness between sentence pairs. Similar to the Stanford Natural Language Inference (SNLI) dataset (Bowman et al., 2015), ASSIN incorporates a label feature that classifies sentence pairs as follows: entailment (0), no apparent relationship (1), or contradiction (2). In the case of ASSIN2, the dataset contains only two labels – 1 and 0 – indicating whether the sentences are entailed or not, respectively.

Adding this type of differentiation, we achieve the following models:

- stjiris/bert-large-portuguese-cased-legal-mlm-nli-sts-v0;[14]
- stjiris/bert-large-portuguese-cased-legal-tsdae-nli-sts-v0.[14]

Generative Pseudo Labeling Generative Pseudo Labeling (GPL) (Wang et al., 2022) is an unsupervised technique designed to improve dense retrievers. The process consists of three stages: Query Generation, Negative Mining, and Pseudo Labeling.

During the Query Generation phase, we employed a pre-trained T5 model, *pierreguillou/t5-base-qa-squad-v1.1-portuguese*,[17] to generate 10,000 queries from legal documents. In the Negative Mining stage, we retrieved similar yet non-informative passages using ElasticSearch and a BERTimbau model fine-tuned for Semantic Textual Similarity (STS). Finally, we computed the margin between positive and negative passages and trained the model utilizing Margin Mean Squared Error Loss. This technique was applied to models that were trained using both NLI and STS data:

- stjiris/bert-large-portuguese-cased-legal-mlm-gpl-nli-sts;[14]
- stjiris/bert-large-portuguese-cased-legal-tsdae-gpl-nli-sts.[14]

6.3 Summarization

In this study, we employed a straightforward, centrality-based summarization technique, *Continuous LexRank*, as it enables a more effective analysis of the impact of domain-adapted embeddings and the segmentation judgments model.

Following the initial pre-processing of the judgments and the identification of summarizable paragraphs in the baseline method described in Section 5, we segmented them into individual sentences using Stanza (Qi et al., 2020).[18] Each summarizable sentence was then processed using the LexRank Python package for summarization.[19] To determine the most informative paragraphs, we aggregated the scores of the sentences within each paragraph and normalized them by the number of sentences. As discussed in Section 4, the selected paragraphs for summarization account for approximately 2% of the total paragraphs in the judgment.

In this configuration, rather than using the baseline pre-processing pipeline described in Section 5, we first segment each judgment into sections with the Segmenting Judgment model and then feed paragraphs from selected sections directly into the LexRank algorithm. We opted not to split these paragraphs into individual sentences for three key reasons: First, in our baseline method, we selected the most relevant paragraphs to minimize the risk of fragmenting an argument by aggregating the scores of individual sentences within them. Second, segmenting paragraphs into sentences using the Stanza package is considerably time-consuming. Finally, our embedding models — trained without such granular separation (with the exception of the Word2Vec model) — demonstrated improved performance when processing entire paragraphs rather than isolated sentences.

As outlined in Section 6, we experimented with multiple configurations in which the summarizer was provided with paragraphs from various sections and produced paragraphs from different sections as output. The specific configurations are as follows:

- Supplying paragraphs exclusively from the **grounds** section to the summarizer and returning paragraphs exclusively from the **grounds** section (g/g).
- Providing paragraphs from the **grounds**, **report**, and **decision** sections to the summarizer and returning paragraphs solely from the **grounds** section (g-r-d/g).
- Feeding paragraphs from the **grounds**, **report**, and **decision** sections to the summarizer and retrieving paragraphs from both the **grounds** and **report** sections (g-r-d/g-r).
- Offering paragraphs from the **grounds** and **report** sections to the summarizer and returning paragraphs exclusively from the **grounds** section (g-r/g).
- Supplying paragraphs from the **grounds** and **report** sections to the summarizer and obtaining paragraphs from both the **grounds** and **report** sections (g-r/g-r).

In each configuration, we conducted experiments with two variants: one that included the consideration of *italic* paragraphs and another that excluded *italic* paragraphs. In line with the configurations described earlier, we explored various embedding models, as detailed in Section 6.2.2. These models are as follows:

- Word2vec[20];
- bert-large-portuguese-cased[21];
- bert-large-portuguese-cased-legal-mlm-sts-v0[22];
- bert-large-portuguese-cased-legal-tsdae-sts-v0[23];
- bert-large-portuguese-cased-legal-tsdae-gpl-v0[24];
- bert-large-portuguese-cased-legal-tsdae[25];
- bert-large-portuguese-cased-legal-mlm-nli-sts-v1[26];
- bert-large-portuguese-cased-legal-tsdae-gpl-nli-sts-v0.[27]

In the case of the Word2Vec model, word embeddings were trained from scratch on the STJ corpus using the skip-gram architecture. The initial word vectors were randomly initialized and subsequently optimized during training. Paragraph embeddings were then obtained by averaging the word vectors in each paragraph and normalizing the resulting representation. For the BERT models, paragraph embeddings are obtained directly. Since some of the judgments used in the summarization experiments also belong to the training set of the Segmenting Judgment model, there is a risk that segmentation errors on those judgments may be underestimated. To assess the impact of this potential

[15] https://huggingface.co/datasets/assin.
[16] https://huggingface.co/datasets/assin2.
[17] https://huggingface.co/pierreguillou/t5-base-qa-squad-v1.1-portuguese.
[18] Version of stanza1.4.1.
[19] https://github.com/crabcamp/lexrank/tree/dev.
[20] https://huggingface.co/stjiris/Word2vec-legal-portuguese.
[21] https://huggingface.co/neuralmind/bert-base-portuguese-cased.
[22] https://huggingface.co/stjiris/bert-large-portuguese-cased-legal-mlm-sts-v0.
[23] https://huggingface.co/stjiris/bert-large-portuguese-cased-legal-tsdae-sts-v0.
[24] https://huggingface.co/rufimelo/bert-large-portuguese-cased-legal-tsdae-gpl-v0.
[25] https://huggingface.co/stjiris/bert-large-portuguese-cased-legal-tsdae.
[26] https://huggingface.co/stjiris/bert-large-portuguese-cased-legal-mlm-nli-sts-v1.
[27] https://huggingface.co/stjiris/bert-large-portuguese-cased-legal-tsdae-gpl-nli-sts-v0.

Table 6
ROUGE recall scores with 95% bootstrap confidence intervals (mean ± half-width) for extractive summarization with LexRank using different embedding models. For each embedding model, we report the best-performing configuration. All domain-adapted models significantly outperform the not fine-tuned *portuguese-cased* baseline ($p < 0.001$, paired bootstrap test, two-sided). For space reasons, the prefix *bert-large* was omitted from the model names.

Embedding model	Optimal configuration	Consider italics?	Divide judgments	R1	Med	R2	Med	RL	Med
Word2Vec	g-r/g	Yes	Segmenting model	45.95 ± 1.55	46.58	20.67 ± 1.32	16.22	43.32 ± 1.42	43.07
portuguese-cased	g-r/g	No	Segmenting model	33.81 ± 1.39	31.51	12.18 ± 1.025	7.35	33.14 ± 1.19	31.84
portuguese-cased-legal-mlm-sts	g-r-d/g	No	Segmenting model	56.38 ± 1.25	57.68	25.29 ± 1.18	22.37	51.31 ± 1.14	51.68
portuguese-cased-legal-tsdae-sts	g-r-d/g	No	Segmenting model	55.73 ± 1.30	57.73	24.63 ± 1.23	21.49	50.7 ± 1.17	51.31
portuguese-cased-legal-tsdae-gpl-nli-sts	g-r/g	No	Segmenting model	**57.23 ± 1.23**	**58.13**	**25.44 ± 1.22**	**22.86**	**51.91 ± 1.11**	**52.17**
portuguese-cased-legal-tsdae-gpl[a]	g-r/g	No	Segmenting model	47.89 ± 1.37	48.87	19.48 ± 1.09	15.79	44.47 ± 1.21	44.44
portuguese-cased-legal-mlm-nli-sts-v1	g-r-d/g	No	Segmenting model	55.94 ± 1.26	57.56	24.88 ± 1.22	21.19	50.77 ± 1.14	50.88

[a] Significantly below all STS fine-tuned variants ($p < 0.001$).

leakage, we ran an additional configuration in which we bypass the segmentation model and use the gold section annotations to partition the judgments. In this configuration, the summarization pipeline operates on the ground-truth sections rather than on the model's predicted sections, while all other components remain unchanged.

When the Segmenting Judgment Model failed to allocate paragraphs to the intended sections for summarization, we chose to replace the respective section with the one containing the highest number of available paragraphs. This approach is grounded in the assumption that the section with the greatest number of paragraphs is more likely to contain the paragraphs we intend to provide to the summarizer.

7 Evaluation

To assess our system, we employ three distinct measures: ROUGE evaluation, assessment using the LLM Llama[28] and GPT-4o,[29] and evaluation conducted by judges from the Portuguese Supreme Court of Justice.

7.1 ROUGE evaluation

The ROUGE metric (Lin, 2004) evaluates the quality of a summary by counting the number of overlapping *n*-grams between the human summary and the automatic summary. In Table 6, we present the results of our experiment, where we use the LexRank technique alongside different embedding models to generate extractive summaries. The relatively high ROUGE scores are consistent with the characteristic of our dataset described in Section 4, namely that judges frequently reuse passages from the judgments when drafting the official summaries.

The current state of the art in this domain is dominated by complex neural models trained directly for summarization, which typically require large collections of judgments paired with gold summaries — a requirement that was not feasible in our setting. Instead, we focus on an unsupervised, extractive approach based on LexRank combined with domain-specific embeddings, providing a strong and interpretable baseline for summarizing Portuguese Supreme Court judgments within the constraints of the available data.

For each embedding model, we selected the optimal configuration based on the ROUGE-1 scores. It is worth noting that, as demonstrated by Lin (2004), ROUGE-1 stands out as the most reliable metric for evaluating summary quality. The respective results and the optimal configurations for each embedding model are presented in Table 6.

The optimal configuration was found to vary solely when employing the Word2Vec embedding model, with better results obtained when considering *italic* paragraphs. Additionally, when using the *bert-large-portuguese-cased-legal-mlm-nli-sts-v1*, the *bert-large-portuguese-cased-legal-tsdae-sts-v0* and the *bert-large-portuguese-cased-legal-mlm-sts-v0* embedding models, the best configuration involved supplying the summarizer with the grounds, report, and decision sections while returning only the paragraphs from the grounds section, instead of feed the summarizer with the grounds and report section. Across all embedding models, the top two results consistently emerge from employing the section distribution of g-r/g and g-r-d/g. For the Word2Vec embedding model, we observe a substantial ROUGE-1 difference of 8.06 points, whereas for the BERTimbau model the difference is 1.13 points. For the remaining models, the differences are negligible, ranging from 0.01 to 0.10 points. When comparing the section distribution between g-r/g and g-r/g-r, the sole differentiation lies in the latter allowing the summarizer to return paragraphs from the report section. In this scenario, the differences in ROUGE-1 scores are more pronounced, ranging from 1.51 to 10.77. The scores of the former consistently outperform those of the latter, and notably, configurations that exclusively return paragraphs from the grounds section consistently yield better results than those returning paragraphs from both the grounds and report sections. These results show that the grounds section is the most important section for the production of the summaries, as discussed in Section 4. Another deduction that can be drawn from the obtained results is the paramount role of context in enhancing the quality of summaries. Consistently, the g/g section distribution yields lower results than the g-r/g and g-r-d/g section distributions. This observation underscores the significance of providing the summarizer with the report and/or decision sections, as it facilitates a more profound understanding of the most critical segments within the grounds section.

Upon analyzing the results presented in Table 6, it is noteworthy that the utilization of the BERTimbau embedding model, which is not specifically tailored to the Portuguese legal domain, yielded inferior results compared to all other models, each of which was adapted for the Portuguese legal domain. This emphasizes the importance of our embedding models having acquired valuable insights within our domain.

Comparing the results of BERTimbau and Word2Vec with those of our baseline method, we observed a decrease in the ROUGE-1 score from 41.98 to 33.81 and from 47.92 to 45.95, respectively (see Tables 6 and 7). Our conclusion that focusing on returning paragraphs from the grounds section leads to superior results suggests that this decline is attributed to the creation of embedding vectors at the paragraph level rather than the sentence level. In particular, this effect can be explained by a mismatch between the training granularity of Word2Vec, which was trained at the sentence level, and its application at the paragraph level in our system. In contrast, Legal-BERT-based models are pretrained on longer textual sequences and applied consistently at the paragraph level, which better aligns with their training regime and mitigates this effect.

However, the remaining embedding models achieved significantly higher results with this new architecture. We obtained an increase of 9.31 in terms of ROUGE-1, representing a 19.4% improvement. Our BERT embedding models perform better at the paragraph level, where the risk of fragmentation is minimized, allowing for a more comprehensive semantic understanding of the legal text. Furthermore, we provide evidence that we can better mitigate the noise present in the judgments by using the Segmenting Judgment Model rather than relying solely on pre-processing the judgments.

[28] https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct.
[29] https://developers.openai.com/api/docs/models/gpt-4o.

Table 7
Rouge metric for extractive summarization (with LexRank) using different embedding models with the rule-based pre-processing pipeline. For size constraints, all instances of bert-large in the models were omitted.

Embedding model	R1	Med	R2	Med	RL	Med
Word2Vec	**47.92**	**62.97**	**22.50**	29.91	**35.89**	**46.82**
portuguese-cased	41.98	58.29	18.52	25.00	31.92	41.98
portuguese-cased-legal-mlm-sts	44.96	61.70	21.11	29.06	34.00	44.54
portuguese-cased-legal-tsdae-sts	43.88	60.05	20.08	28.07	33.14	43.20
portuguese-cased-legal-tsdae-gpl-nli-sts	46.00	62.35	21.85	**30.00**	34.83	45.47
portuguese-cased-legal-tsdae-gpl	46.74	62.90	21.78	29.85	35.04	45.77

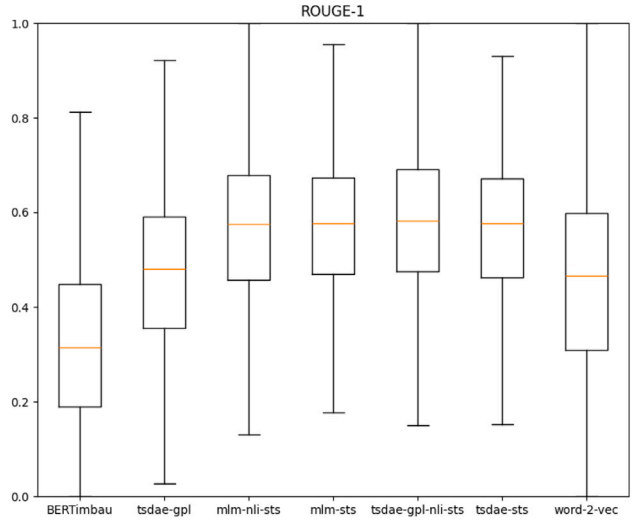


Fig. 9. ROUGE-1 box plot for all models.

The embedding model *bert-large-portuguese-cased-legal-tsdae-gpl-nli-sts-v0* achieved the highest scores across all metrics. However, the more informative observation is the pattern of results across the legal-domain variants rather than the identity of the top performer. The four variants that include STS fine-tuning — *mlm-sts*, *tsdae-sts*, *mlm-nli-sts-v1*, and *tsdae-gpl-nli-sts* — cluster tightly within a 1.5-point ROUGE-1 range (55.73 to 57.23), regardless of which upstream training stages they combine (MLM, TSDAE, GPL, NLI). In contrast, the single legal-domain variant without STS fine-tuning, *tsdae-gpl*, drops to 47.89 ROUGE-1, an absolute decrease of roughly 8 points relative to its STS-trained counterparts. This pattern is mechanistically consistent with the role similarity plays in LexRank's scoring function: LexRank computes paragraph centrality from pairwise cosine similarities, so an embedding space that has been explicitly optimized for similarity judgments is precisely what LexRank exploits at inference time. STS fine-tuning therefore emerges as the single training component whose presence or absence visibly affects summarization performance, while the remaining upstream stages contribute marginally on this task. The additional exposure to Natural Language Inference (NLI) data is consistent with the slight edge of NLI-trained variants, in line with the findings reported by Reimers and Gurevych (2019) that NLI supervision improves STS performance, but the residual differences among the STS-trained variants (under 1.5 ROUGE-1 points) are small enough that we do not attempt to rank them against each other. We treat the four STS-trained variants as performance-equivalent for this task, and the per-component contributions of TSDAE, GPL, and NLI are analyzed in detail in the work by Melo et al. (2023). Taken together, these results underscore the importance of domain adaptation: the best legal-domain variant achieves a ROUGE-1 score of 57.23, compared to 33.81 for the unadapted BERTimbau baseline — an absolute improvement of 23.42 points — confirming that domain-specific pretraining is the dominant factor driving performance on this task.

The exclusion of *italic* paragraphs consistently resulted in improved performance across all models, except for the Word2Vec model. This outcome aligns with our expectations, as *italic* paragraphs, which usually consist of citations as elucidated in Section 6, are not used in the summary generation process. Surprisingly, refraining from using the ground-truth dataset for segmenting the judgments resulted in slightly improved performance across all models. The judges' lack of consistency when writing judgments, with proven facts sometimes found in the report section and other times in the grounds section, suggests that our model may be more adept at capturing the overall structure in a more coherent manner. As elucidated in the preceding section, in cases where the segmenting model failed to assign any paragraph to the grounds section, our approach involved selecting the section with the highest number of paragraphs. To assess the performance of our best configuration, we examined the origins of the paragraphs. Out of the 3028 paragraphs in total, 3017 belonged to the grounds section, 1 to the footnote section, 9 to the report section, and 1 to the decision section. Out of the 700 judgments, in 99.6% of instances, the model successfully assigned paragraphs to the grounds section.

To strengthen the statistical validity of the reported results, we computed 95% bootstrap confidence intervals (10,000 resamples) and paired bootstrap significance tests (two-sided, 10,000 resamples) for all pairs of embedding models, reported in Table 6. The confidence intervals are narrow — ranging from ±1.03 to ±1.55 points across all models and metrics — indicating that the reported scores are stable estimates rather than artifacts of the particular sample of judgments in the test set. The significance tests confirm that STS fine-tuning is the decisive training component: the gap between the non-STS legal variant (*tsdae-gpl*) and all four STS-trained variants is highly significant across all metrics ($p < 0.001$), while differences among the four STS-trained variants, though occasionally reaching nominal significance for ROUGE-1 and ROUGE-L, remain below 2 points in absolute terms and are unlikely to be of practical relevance.

Fig. 9 presents the distribution of ROUGE-1 scores across all models for the entire corpus. Some extractive summaries achieved a ROUGE-1 score very close to 1, indicating that nearly every word in the automatic summary was present in the human-generated summary.

For instance, Table 8 compares an extractive summary with its corresponding human-generated summary for the same judgment. In terms of the ROUGE-1 and ROUGE-L metrics, the automatically generated summary attained scores of 0.98 and 0.96, respectively. The words highlighted in bold are those that appear in both summaries. Notably, our architecture successfully retrieved the same paragraphs from the judgment that were used to construct the human-generated summary, demonstrating its effectiveness in identifying the most critical paragraphs within the judgment.

On the other hand, certain automatically generated summaries achieve ROUGE-1 and ROUGE-L scores of 0 or close to 0, as exemplified in Table 9. This case clearly demonstrates that our system failed to extract the critical paragraphs necessary for generating a meaningful summary. All of these paragraphs consist of factual information that does not contribute to the understanding of the case or its verdict.

One potential approach to mitigate this issue is to further segment the *grounds* section into distinct sub-sections, such as *facts* and *reasoning*, and exclusively return paragraphs from the *reasoning* sub-section. The details of these potential improvements are further discussed in Section 9.

Table 8
Comparison between the human-generated summary and a good automatic-generated summary example. Both summaries were translated into English while preserving identical translations for analogous words.

Human-summary	Automatic-summary
- In terms of Article 854 of the Civil Procedure Code, "without prejudice to cases where an appeal to the Supreme Court of Justice is always admissible, a review is only possible, in general terms, of the judgments of the Court of Appeal issued in appeals in liquidation procedures not dependent on simple arithmetic calculation, verification, and ranking of credits, and opposition filed against execution". -A contrary interpretation of the norm leads to the conclusion that a review of decisions regarding the main execution instance is not admissible, but only of decisions regarding its declaratory components.	We are dealing with an execution action in which the review is provided for in Article 854 of the Civil Procedure Code, in terms of which, "without prejudice to cases where an appeal to the Supreme Court of Justice is always admissible, a review is only possible, in general terms, of the judgments of the Court of Appeal issued in appeals in liquidation procedures not dependent on simple arithmetic calculation, verification, and ranking of credits, and opposition filed against execution". A contrary interpretation of the norm leads to the conclusion that a review of decisions regarding the main execution instance is not admissible, but only of decisions regarding its declaratory components.

Table 9
Comparison between the human-generated summary and a bad automatic-generated summary example.

Human-summary	Automatic-summary
I. The principle of preclusion established in Article 573 of the CPC refers to the assertion of relevant facts for triggering the means of defense and the expression of the will to activate those means of defense that the law makes dependent on the will of the party concerned, but not yet to the invocation of applicable legislation and the resulting consequences. II. The violation of the principle of adversarial proceedings under Article 3, paragraph 3, of the CPC does not give rise to a procedural nullity under Article 195 of the CPC, but rather to a nullity due to excess of pronouncement, pursuant to Articles 615, paragraph 1, subparagraph (d), 666, paragraph 1, and 685 of the same statute. III. The principle of adversarial proceedings established in Article 3, number 3, of the CPC does not imply systematic and trivial recourse, perhaps even dilatory, to the hearing of the parties. It should only take place if the parties have not had the opportunity to express themselves and if there is no clear case of unnecessary proceedings. IV. bullet 11 of Annex I to Regional Legislative Decree 7/2006/M (previously bullet 8 of Annex II to Regional Legislative Decree 2/2204/M) establishes a minimum duration period for full maintenance contracts, unless a shorter period is expressly stipulated, without preventing the establishment of longer contractual duration periods. V. A penalty clause amounting to 50% of the remunerations due until the end of the contract appears to be manifestly excessive in the circumstance of a contract with an 18-year duration terminated in the fourth month of its existence, and should be reduced by 75%.	Number FCN16003125, with a payment deadline on 01.05.2016, in the amount of €113,09. Number FCC13004917, with a payment deadline on 01.08.2013, in the amount of €564.59; Number FCC13006928, with a payment deadline on 01.11.2013, in the amount of €564.59; Number FCC14001976, with a payment deadline on 01.03.2014, in the amount of €366.00; Number FCC14005002, with a payment deadline on 01.07.2014, in the amount of €366.00; Number FCC14008715, with a payment deadline on 01.12.2014, in the amount of €376.98; Number FCC16003012, with a payment deadline on 01.04.2016, in the amount of €376.98; Number FRZ16002074, with a payment deadline on 19.12.2016, in the amount of €20.01.

7.1.1 Error propagation analysis

In Section 6.1.3, we identified, among the 134 original test judgments, the 18 judgments whose segmentation errors directly affect the Report or Grounds sections — the sections that feed the summarizer — and deferred to this section the question of whether those errors translate into measurable degradation of summary quality. A first, aggregate-level answer is already available from the comparison reported earlier in this section: replacing the predicted segmentation with gold section annotations does not improve summarization performance, and in fact slightly degrades it across all models. As discussed above, this counter-intuitive result is plausibly explained by the inconsistency with which judges themselves organize judgments — proven facts, for instance, appear sometimes in the Report and sometimes in the Grounds section — so that the model's predictions may capture the functional structure more coherently than the reference annotations. This already suggests that the pipeline is, in aggregate, robust to the errors the segmentation model makes. We now complement that aggregate view with a finer-grained analysis, isolating the specific judgments affected by segmentation errors to verify that the conclusion holds even when attention is restricted to the cases most at risk.

Building on the document-level error analysis of Section 6.1.3, we partition the 134 original test judgments into three groups: those with no segmentation errors (102 judgments), those whose errors do not affect the sections relevant to summarization (14 judgments, with only intra-section confusions or inter-section errors between non-summarization sections), and those with errors that directly affect the Report or Grounds sections (18 judgments). For each group, we compute the ROUGE recall of the generated summaries against the human references, using our best configuration (*bert-large-portuguese-cased-legal-tsdae-gpl-nli-sts*). Table 10 reports the average ROUGE recall per group, and Fig. 10 shows the corresponding distributions. The three groups are remarkably similar on ROUGE-1 and ROUGE-L recall, with no consistent ordering and differences within roughly two points — the judgments with summarization-relevant errors do not score systematically lower than the error-free group. On ROUGE-2 the group with summarization-relevant errors shows a somewhat lower median, but the distributions still overlap substantially and the within-group variance dwarfs the between-group gap (see Fig. 10); with only 18 affected judgments, this difference is well within the noise expected at this sample size.

This absence of measurable degradation is consistent with the structure of the errors themselves. As characterized in Section 6.1.3, the dominant summarization-relevant error is the inclusion of *spurious* paragraphs rather than the loss of relevant ones, by a ratio of more than four to one. Spurious paragraphs enlarge the candidate pool passed to LexRank but do not remove genuine content. The nature of the dominant spurious source — Footnote paragraphs misclassified as Grounds — is particularly relevant here: as discussed in Section 6.1.1, footnotes in STJ judgments frequently contain substantive legal citations, clarifications, and argumentative remarks that closely mirror the reasoning found in the Grounds section. Rather than introducing off-topic noise, these paragraphs may in fact reinforce the central argumentative cluster that LexRank identifies, potentially contributing positively to the centrality scores of genuinely relevant paragraphs. The secondary spurious source, Report paragraphs misclassified as Grounds, is likewise intended for summarization, and its inclusion is

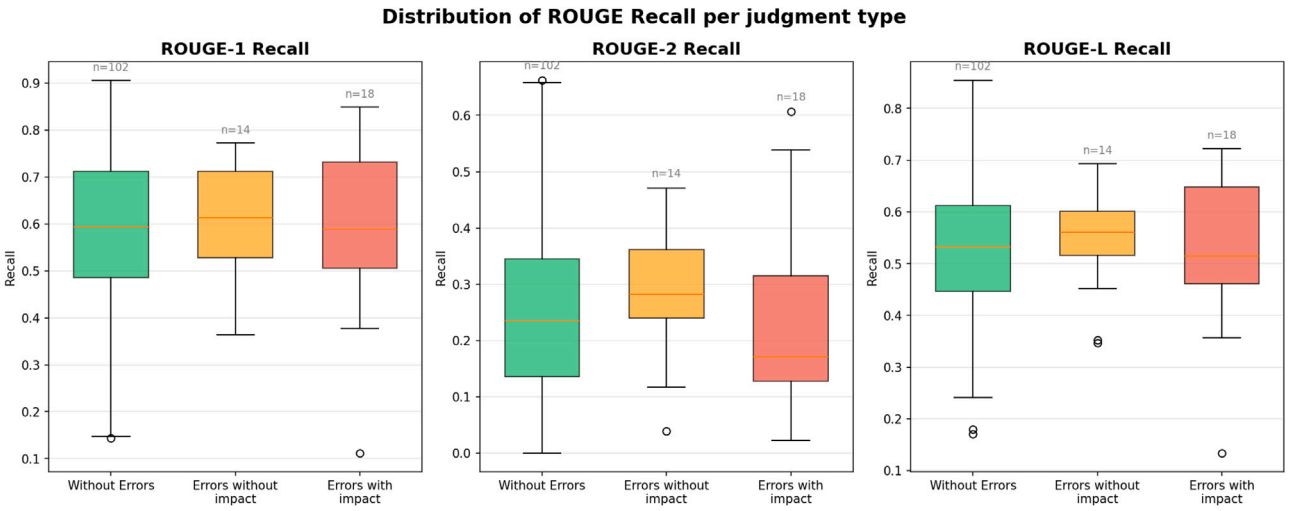


Fig. 10. Distribution of ROUGE recall (ROUGE-1, ROUGE-2, ROUGE-L) of the generated summaries across the three judgment groups defined by segmentation error type. The number of judgments per group is indicated above each box.

Table 10
Average ROUGE recall of the generated summaries (best model: *bert-large-tsdae-gpl-nli-sts-v0*) for the 134 original test judgments, grouped by the type of segmentation error they contain. *Errors without impact* denotes judgments whose errors affect only sections irrelevant to summarization; *errors with impact* denotes judgments with at least one inter-section error involving the Report or Grounds sections.

Judgment group	N	ROUGE-1	ROUGE-2	ROUGE-L
Without errors	102	0.5798	0.2569	0.5257
Errors without impact	14	0.5987	0.2901	0.5501
Errors with impact	18	0.5908	0.2334	0.5183

therefore unlikely to degrade the output. For the comparatively rare *lost* paragraphs — the more dangerous failure mode — the affected judgments are too few and the losses not severe enough to register at the group level. We therefore find no evidence that segmentation errors, at the rate and of the type produced by our model, propagate into measurable summarization degradation — though we note that this conclusion is bounded by the size of the test set, and that a larger affected sample might expose effects too small to detect here.

7.2 LLMs' evaluation

To further assess the quality of our extractive summaries, we employed two state-of-the-art LLMs to evaluate them across three dimensions: *Coherence*, *Completeness*, and *Accuracy*. This evaluation was designed as a complementary perspective, capturing semantic and structural dimensions of quality that lexical overlap metrics are not equipped to measure. Additionally, we analyzed the correlation between the LLM evaluations and the ROUGE scores to identify where the two evaluation approaches converge and, crucially, where they diverge.

7.2.1 Llama evaluation

We used the Llama 3.1 8B instruct LLM to evaluate our automatic summaries in three different parameters: Coherence, Completeness and Accuracy. Fig. 11 shows the prompt we used for this evaluation.

Inference was run deterministically with greedy decoding (*do_sample=False*), and *max_new_tokens = 512*. Each summary was evaluated once (one inference call per summary) under the same settings to ensure reproducibility.

In terms of the Accuracy and Coherence parameters, the results are highly satisfactory, with a average of 4.93 for accuracy and 4.53 for Coherence. However, Completeness attained a average of 3.76, indicating

You are a jurisprudence extractive summary evaluator. Given a <jurisprudence ></jurisprudence> and a <summary></summary>, your task is to grade the summary in the following three dimensions, from 1 to 5:
Coherence - This evaluates how logical and well-structured the summary is. A good summary should flow properly from one idea to the next. Extracted sentences should connect well without seeming disjointed.
Completeness - This measures whether the most essential information and key points from the original document are included in the summary. It is important that the main concepts and ideas are captured in the extraction process.
Accuracy - This verifies that the content in the summary factually represents what is contained in the original document without inserting incorrect or false information. The reliability and faithfulness of the summary should be assessed. Any factual inconsistencies or errors would make the summary less useful.
Output your answer like so:
Coherence: score
Completeness: score
Accuracy: score

Here is the jurisprudence document:
<jurisprudence>
jurisprudence
</jurisprudence>

Here is the context:
<summary>
summary
</summary>

Fig. 11. Llama Prompt to evaluate summaries across three dimensions.

a comparatively lower performance, as we do not consider paragraphs from the Report Section (resulting in a lack of completeness), following the approach adopted by the judges.

Among the three dimensions, Completeness is conceptually the closest to what ROUGE measures, as both are sensitive to whether key content from the source is present in the summary. However, as shown in Fig. 12, the association between ROUGE-1 and the LLM-based completeness scores is weak and highly dispersed. For completeness

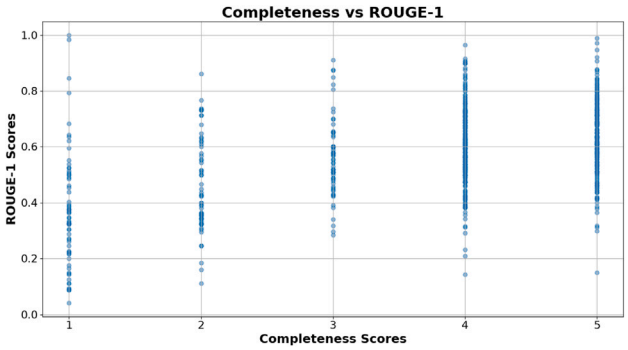


Fig. 12. Scatterplot showing the relationship between Llama-assigned completeness scores and ROUGE-1 scores for the generated summaries. Each column corresponds to a completeness rating, with ROUGE-1 values distributed vertically.

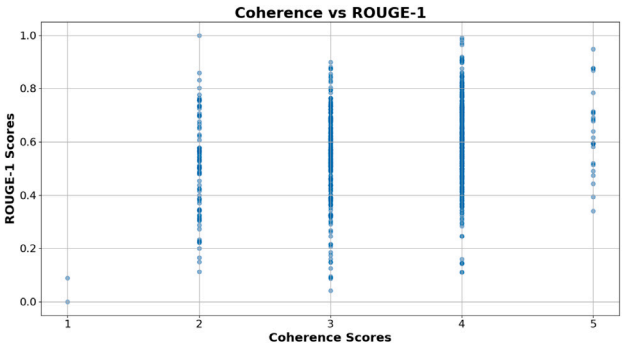


Fig. 14. Scatterplot showing the relationship between GPT 4o-assigned coherence scores and ROUGE-1 scores for the generated summaries. Each column corresponds to a coherence rating, with ROUGE-1 values distributed vertically.

System Message:
OpenAI system: You are an extractive summary evaluator. Given a <document > </document >and a <summary ></summary >, your task is to grade the summary in the following three dimensions, from 1 to 5:
Coherence - This evaluates how logical and well-structured the summary is. A good summary should flow properly from one idea to the next. Extracted sentences should connect well without seeming disjointed.
Completeness - This measures whether the most essential information and key points from the original document are included in the summary. It's important that the main concepts and ideas are captured in the extraction process.
Accuracy - This verifies that the content in the summary factually represents what is contained in the original document without inserting incorrect or false information. The reliability and faithfulness of the summary should be assessed. Any factual inconsistencies or errors would make the summary less useful.
Output your answer like so:
Coherence: score
Completeness: score
Accuracy: score

OpenAI user: Here is the document:

<document> document </document>

Here is the context:
<summary> summary </summary>

Fig. 13. GPT-4o Prompt to evaluate summaries across three dimensions.

score 1, ROUGE-1 spans nearly the full range, with a higher concentration of values below 0.4. For score 2, ROUGE-1 is concentrated mainly between 0.3 and 0.7, whereas for score 3 it falls mostly between 0.4 and 0.6. For scores 4 and 5, ROUGE-1 values are predominantly above 0.4, mixing medium and high lexical overlap. Overall, these patterns indicate that lexical overlap captures only a surface proxy of Completeness, and that the two measures are tapping into related but distinct aspects of summary quality. The divergence is even more pronounced for Coherence and Accuracy, where no structural parallel with ROUGE exists, and the scores reflect dimensions — logical flow and factual fidelity — that lexical overlap is fundamentally unable to assess.

7.2.2 GPT-4o evaluation

We also used GPT-4o to evaluate our automatic summaries in the same three different parameters: Coherence, Completeness and Accuracy. Fig. 13 shows the prompt we used for this evaluation. For GPT-4o, we queried the model through our institutional API endpoint (IAEDU[30]), providing the evaluation prompt as the request message. In our current setup, decoding parameters are not explicitly set in the client code and therefore follow the service defaults. Each summary was evaluated once (one API request per summary).

Comparatively, when contrasted with Llama LLM, the outcomes derived from GPT-4o demonstrate a reduction. In terms of accuracy, GPT-4o attained an average of 4.43, reflecting a decrease of 0.5. Regarding coherence, GPT-4o achieved an average of 3.48, indicating a decrease of 1.05. For the completeness parameter, GPT-4o yielded an average of 2.81, marking a decrease of 0.95. The diminished completeness is attributed to the same factor elucidated earlier, the omission of paragraphs from the Report section, leading to an absence of completeness.

Nevertheless, GPT-4o's scores show a slightly stronger association with ROUGE-1 than Llama scores. The pattern is clearest at score 5, where ROUGE-1 values tend to be higher, although some dispersion remains (Fig. 14). Score 1 contains only two observations, which is insufficient for meaningful interpretation. For scores 2–4, ROUGE-1 values are widely spread and strongly overlapping across score levels, indicating that ROUGE-1 provides limited discrimination between coherence ratings. Fig. 15 indicates a weak but noticeable upward trend between GPT-4o completeness scores and ROUGE-1: as completeness increases, higher ROUGE-1 values become more frequent, particularly at scores 4–5. However, ROUGE-1 remains widely dispersed within each score and strongly overlapping across adjacent levels, suggesting that lexical overlap provides limited discrimination of completeness. For score 3, ROUGE-1 values are concentrated mostly in the mid-range (approximately 0.4–0.8), consistent with moderately complete summaries exhibiting moderate overlap.

7.3 Evaluation conducted by judges from the portuguese supreme court of justice

To conduct this expert evaluation, we convened with two Judges in a panel-based discussion from the Portuguese Supreme Court of Justice. Unfortunately, achieving an odd number of judges, as is customary in this type of evaluation, proved unattainable. Nevertheless, given that this evaluation did not adhere to a systematic approach, wherein judges are required to assign scores to each summary, but rather followed an open evaluation format, the insights gleaned from the assessment conducted with two judges proved to be substantial.

Table 11
ROUGE and LLM-based evaluation scores (Llama and GPT-4o) for the five best approach summaries assessed by the judges. Ch = Coherence; Cp = Completeness; Ac = Accuracy.

ID	ROUGE			Llama			GPT-4o		
	R1	R2	RL	Ch	Cp	Ac	Ch	Cp	Ac
Ac. 3421-16.3T8FNC.L1.S1	0.98	0.97	0.96	2	1	2	4	2	5
Ac. 170.19.4TYHLSB.L1.S1	0.4	0.3	0.45	5	2	5	2	1	4
Ac. 1012-15.5T8VRL-AU.G1-A.S1	0.8	0.36	0.59	4	3	5	4	3	5
Ac. 12131-18.6T8LSB.L1.S1	0.27	0.05	0.19	2	1	5	3	2	4
Ac. 19722.18.3T8LSB.L1.S1	0.15	0.01	0.15	2	1	5	4	2	5

Table 12
First summary example.

Ac. 3421-16.3T8FNC.L1.S1		
Human-summary	Best approach summary	Prior work summary
- Under the provisions of Article 854 of the Code of Civil Procedure, "without prejudice to cases where an appeal to the Supreme Court of Justice is always admissible, an appeal is only permitted, under the general terms, against decisions of the Court of Appeal rendered in proceedings concerning liquidation not dependent on simple arithmetic calculation, verification and ranking of claims, and opposition filed against enforcement". - An *a contrario* interpretation of the provision leads to the conclusion that an appeal is not admissible for decisions pertaining to the main enforcement proceedings, but only for decisions related to their declaratory offshoots.	We are dealing with an enforcement action in which an appeal to the Supreme Court is provided for under Article 854 of the Code of Civil Procedure, according to which, "without prejudice to cases where an appeal to the Supreme Court of Justice is always admissible, an appeal is only permitted, under the general terms, against decisions of the Court of Appeal rendered in proceedings concerning liquidation not dependent on simple arithmetic calculation, verification and ranking of claims, and opposition filed against enforcement". An *a contrario* interpretation of the provision leads to the conclusion that an appeal is not admissible for decisions pertaining to the main enforcement proceedings, but only for decisions related to their declaratory offshoots.	We are dealing with an enforcement action in which the appeal process is provided for under Article 854 of the Code of Civil Procedure, which states that, "without prejudice to cases where an appeal to the Supreme Court of Justice is always admissible, an appeal is only permitted, under the general terms, against decisions of the Court of Appeal rendered in proceedings concerning liquidation not dependent on simple arithmetic calculation, verification and ranking of claims, and opposition filed against enforcement". "As a rule, the appealability of the Court of Appeal's judgments on the opposition filed against the attachment and on the generality of interlocutory decisions is excluded, whether challenged together with the appeal of the final decision or independently".

The ROUGE-1, ROUGE-2 and ROUGE-L scores for the five summaries evaluated by the judges are reported in Table 11, together with the corresponding Llama and GPT-4o coherence and completeness ratings. These examples were selected to cover a range of performance levels, including one near-perfect case and others exhibiting moderate to low lexical overlap with the human summaries. Given the small number of examples and judges, this evaluation cannot coverage of all data patterns, but can provide good qualitative insights.

The first example illustrates a case in which our system produces a near-verbatim match to the human summary (ROUGE-1 = 0.98, ROUGE-2 = 0.97, ROUGE-L = 0.96). Despite this exceptionally high lexical correspondence, Llama assigns relatively low ratings for coherence, completeness, and accuracy, whereas GPT-4o evaluates the same summary more favorably (except for completeness). This divergence highlights a key limitation of LLM-based evaluation: different models can yield substantially different judgments, even for summaries that closely match the reference.

For each example, we asked each judge to freely comment on each summary. Regarding our first example (see Table 12), our baseline method's summary, the judges criticized it as a poor example due to the inadequacy of the second paragraph, which fails to capture crucial information from the judgment. In contrast, the summary from our current best approach received positive feedback, aligning with its similarity to the original summary. However, a critique was raised against the first sentence, "We are dealing with an enforcement action in which the appeal process is provided for under Article 854 of the Code of Civil Procedure", highlighting that such a sentence is unsuitable for inclusion in a summary. This issue is recognized as a common challenge in extractive summarization, emphasizing the need to reorganize extracted sentences to create more readable and summary-appropriate outputs. One judge expressed a dislike for including article numbers in the summary. The process of creating a summary is indeed highly subjective, and, as evident from these examples, the original summaries contain numerous article numbers. This highlights the inherent personal nature of summary creation, wherein individual preferences and perspectives play a significant role in shaping the content and style of the summaries. As an aid to the judges, the second summary is deemed exemplary as it encapsulates the essential information from the judgment. The only aspect requiring attention is the reorganization of the text to craft a coherent summary.

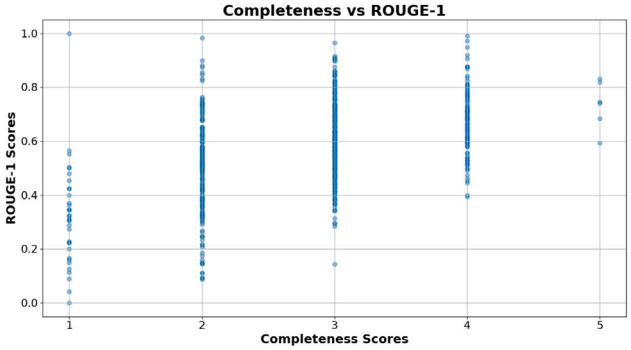


Fig. 15. Scatterplot showing the relationship between GPT 4-assigned completeness scores and ROUGE-1 scores for the generated summaries. Each column corresponds to a completeness rating, with ROUGE-1 values distributed vertically.

Regarding the second example presented in Table 13, it is noted that the best approach summary should exclude the sentence "Thus, it is concluded that". Similar to the first example, there is an issue with sentence connectors. Additionally, the best approach summary requires the inclusion of the first paragraph from the human-generated summary to achieve completeness. Our methodology for determining the number

[30] https://iaedu.pt/pt.

Table 13
Second summary example.

Ac. 170.19.4TYHLSB.L1.S1		
Human-summary	Best approach summary	Prior work summary
I. The conformity of judgments, under the terms of Article 671, paragraph 3, of the Code of Civil Procedure, only allows for an exceptional appeal when such conformity of judgments constitutes an obstacle to the filing of the appeal. II. Since the contradiction of judgments is not decisive, an appeal based on the provisions of paragraph 2, item (d) of Article 629 of the Code of Civil Procedure is not admissible.	Thus, it is concluded that, since the contradiction of judgments is not decisive, the appeal filed does not fall within the scope of the provisions of paragraph 2, item (d) of Article 629 of the Code of Civil Procedure.	Although the provision refers to contradiction with another decision of the Court of Appeal, it has been understood that, a fortiori, the same applies to decisions of the Supreme Court of Justice (ABRANTES GERALDES, PAULO PIMENTA, and L. F. PIRES DE SOUSA, *Código de Processo Civil Anotado*, I, 2018, p. 753).

Table 14
Third summary example.

Ac. 1012-15.5T8VRL-AU.G1-A.S1	
Human-summary	Best approach summary
A judgment cannot serve as the basis for an extraordinary appeal for revision, as it cannot be classified as a document, for the purposes of the provisions of Article 696, item (c), of the Code of Civil Procedure.	Expunging the conclusions formulated by the appellants of what merely constitutes argumentation or is irrelevant for the decision to be made, we can conclude that the issue at hand, within the scope of the appeal, is to assess whether the request for an extraordinary appeal for revision includes the appeal grounds set forth in item (c) of Article 696 of the Code of Civil Procedure invoked by the applicants. The ground invoked in the present case was that of item (c) of Article 696 of the Code of Civil Procedure, according to which a final decision can only be subject to revision when a document is presented that the party did not have knowledge of, or could not make use of, in the process in which the decision to be revised was rendered, and that, by itself, is sufficient to alter the decision in a manner more favorable to the losing party. In conclusion, as stated in the singular decision of 19.04.2020, "the ground of Article 696(1)(c) of the CPC does not contemplate any possibility of an extraordinary appeal for revision to review and reappraise the same evidence already produced and assessed in the first instance, and already reappraised in the appeal by the Court of Appeal, in which a challenge to the facts was made, even if a decision in a different direction has been issued in a case with similar circumstances and evidence". The rendering of a judgment or a decision, even in a cited similar case and even if different criteria were applied to the analysis of evidence and legal appreciation of the facts in this case, does not constitute any other ground for an appeal for revision, among the other provisions of the law in Article 696 of the CPC.

of paragraphs returned for a judgment involves providing 2% of the total paragraphs in the respective judgment. This clarifies why this summary only consists of one paragraph.

It is worth mentioning that there is a contradiction in the judges' feedback. While they suggested extracting 3 to 5 paragraphs as the optimal approach, they also deemed the summary in the third example of our best approach (see Table 14), containing four paragraphs, as too lengthy.

In the summary of our baseline method, the text within the parentheses should be excluded from the summary. However, the information preceding that section is deemed valuable for the summary, albeit requiring reorganization for improved coherence.

Regarding the third example presented in Table 14, the judges emphasized that the human-generated summary serves as an exemplar of a good summary. They praised its conciseness and noted that it encapsulates the crucial aspects of the respective judgment. The summary produced by our best approach for the third example is deemed excessively lengthy, with only the last two paragraphs contributing to the essence of the original summary. As observed in prior examples, these sentences require restructuring to formulate a coherent summary. As an auxiliary tool, our best approach summary fulfills its purpose.

For our fourth example, presented in Table 15, it is considered a poor illustration of a summary. The paragraph structure deviates from correctness, similar to the issues observed in previous examples, and the content significantly differs from the human-generated summary.

Concerning the fifth example presented in Table 16, the summary of our baseline method is deemed a poor example, with content unrelated to the human-generated summary. Additionally, the paragraph in our best approach summary is part of the report section and does not contribute meaningfully to the human-generated summary. The paragraphs generated by humans are sourced primarily from the law sub-section, except the first paragraph that pertains to the facts section. It is noted that the content from the facts section is deemed nonessential for the summary. The second and third paragraphs contain crucial information about the respective judgment.

Based on the judges' feedback on the five examples provided, it appears that there is a correlation between the utility of an automatic summary as an auxiliary tool for judges and the ROUGE-1 score. Higher ROUGE-1 scores seem to indicate more useful summaries for the judges, while lower ROUGE-1 scores are associated with summaries that are deemed less useful by the judges. This correlation suggests that the ROUGE-1 metric may be a reliable indicator of the effectiveness of automatic summaries in aiding judges.

While higher ROUGE-1 scores tend to align with the judges' perception of summary utility in this small sample, the cross-evaluation analysis presented in Section 7.4 shows that this alignment does not hold uniformly at the individual summary level. We therefore treat ROUGE as a useful but insufficient evaluation signal. Nevertheless, given the highly extractive nature of STJ summaries — where judges themselves select paragraphs directly from judgments — optimizing for ROUGE remains a meaningful objective for improving the summarization system: a model that better identifies the critical paragraphs will, by construction, produce higher lexical overlap with the human reference. We revisit this point in Section 9. It is also part of our future plans to train a model for each category of the legal procedure. These categories include eight sections: (1) Civil, (2) Civil, (3) Criminal, (4) Social, (5) Criminal, (6) Civil, (7) Civil, and litigation, For further details on this legal division, see Zanatti, Ribeiro, Pinto, Borbinha (2025). This targeted approach aims to enhance the model's performance by

Table 15
Fourth summary example.

Ac. 12131-18.6T8LSB.L1.S1	
Human-summary	Best approach summary
I – The nullity due to failure to address a matter, which represents the legal sanction for the violation of the provisions in paragraph 2 of Article 608 of the CPC, only occurs when the judge fails to rule on the "issues" submitted for their examination by the parties, or that they should address ex officio, with such issues being considered as the claims made by the parties, but not the arguments invoked, nor the mere legal qualification provided by the litigants.	As clearly evidenced by the content of the decision under review, all these issues were assessed and substantiated in light of the provisions deemed applicable, concluding that there is no basis to consider that, in the present case, the defendant had waived the invocation of legally admissible defenses against the original creditor, particularly the defense of non-performance of the contract.

Table 16
Fifth summary example.

Ac. 19722.18.3T8LSB.L1.S1		
Human-summary	Best approach summary	Prior work summary
1- In the appeal, it is only necessary to analyze whether the court below made a correct application of the regime in Article 662, by altering the facts in accordance with the requests made to it – and these requests are those of the appellant (A.) and not those of the respondent (R.), who is the current appellant, as appeals are limited by the conclusions of the appellants' briefs (the respondent in the appeal was not the appellant) and by matters to be examined ex officio. 2- In an action for the rendering of accounts, where it is not possible to separate the knowledge held by the plaintiff in her capacity as head of the family and as an heir, in light of the facts established, the demand made by the plaintiff for the defendant to render accounts, after all these years, is unjustifiable. It is manifestly unreasonable to require that the defendant be in a position to explain or demonstrate the full exercise of functions, for which he no longer has the necessary information, as it is now in the possession of the A. 3- The demand made by the A. appears manifestly abusive on her part, due to the provisions of Article 334 of the Civil Code, which is why it should be understood that the defendant is no longer obligated to render the accounts for the period in which he was head of the family.	The decision cited by the claimant argued that there is only one level of appeal, explicitly addressing the issue. The challenged decision admitted the appeal without a direct and express ruling on the existence of one or more levels of appeal, which in itself also means that we are not dealing with a true contradiction of jurisprudence, as defined by this Supreme Court of Justice.	6. Regarding the alleged lateness of the request for nullities, it should be noted that it is not valid: the notification of the challenged decision was on March 10, 2021 (cf. Articles 138 and 248 of the CPC), with the 10-day period expiring on March 25, and the 3 additional business days (penalty under Article 139) ending on March 30, the date the request was filed (the last day of the penalty — which was paid).

tailoring it to the specific characteristics and nuances of each legal category, leading to the generation of different summaries for each category. The judges concurred with this approach but emphasized that the thematic content of the judgment holds greater significance in determining the type of summary produced. For instance, in judgments related to road accidents, it is crucial to include details such as the age of the person who suffered the accident and the compensation received. This information becomes instrumental in similar cases, where individuals of similar ages who have experienced a road accident may receive comparable compensation. The thematic relevance thus play a pivotal role in tailoring the summaries to address specific and pertinent details for each legal scenario.

Some summaries produced by the judges incorporate jurisprudence. We inquired about the criteria employed to ascertain when a judgment incorporates a summary that comprises jurisprudence. The judges explained that this inclusion occurs when there are conflicting decisions for similar cases. Crafting a summary of jurisprudence provides a clear path for subsequent similar cases to follow, ensuring consistency in legal interpretations and decisions. These summaries, which include jurisprudence, are situated in the decision section. Consequently, the creation of such summaries differs from the process followed for other types of summaries. The specific placement within the decision section highlights the unique nature and purpose of jurisprudential summaries in providing a clear reference for future cases. These types of summaries are beyond the scope of this work.

7.4 Cross-evaluation analysis: Reconciling metric disagreement

The three evaluation approaches reported in the previous subsections — ROUGE, LLM-based scoring, and qualitative feedback from STJ judges — provide complementary perspectives on summary quality, but they do not converge. In this subsection we examine where the three signals agree and, more importantly, where they diverge, and we discuss what this disagreement reveals about the current state of summarization evaluation in the legal domain.

A first source of disagreement lies between ROUGE and the LLM-based scores. The scatterplots in Figs. 12, 14 and 15 show that, even for the dimension most conceptually aligned with lexical overlap — Completeness — ROUGE-1 values are widely dispersed within each LLM score level, with substantial overlap across adjacent levels. A second source of disagreement lies between the two LLMs themselves. As reported in Sections 7.2.1 and 7.2.2, GPT-4o produces systematically lower averages than Llama and the disagreement is not only one of magnitude. Table 11 contains striking instance-level inversions: for the near-verbatim summary Ac. 3421-16.3T8FNC.L1.S1, with R1 = 0.98, Llama assigns 2/1/2 across the three dimensions, while GPT-4o assigns

4/2/5. A third source of disagreement lies between automatic metrics and the judges. The same Ac. 3421-16.3T8FNC.L1.S1 example, despite its near-perfect lexical overlap with the human reference, was criticized by the judges because of an opening sentence deemed inadequate for a legal summary — a defect that is invisible to ROUGE and was not captured uniformly by the LLM evaluators.

We identify four structural reasons for the observed inconsistencies, all of which are particularly pronounced in the legal domain. First, lexical overlap is satisfied by formulaic language: legal summaries rely heavily on recurring article references, procedural formulas, and court designations, which inflate ROUGE-1 scores even when the substantive content of the summary is misaligned with the reference. This is consistent with the finding of Zanatti, Ribeiro, Pinto (2025), who report that ROUGE-2, by capturing bigrams rather than isolated tokens, correlates more strongly with expert judgments than ROUGE-1 across all evaluated dimensions on Portuguese Supreme Court summaries. Second, LLM-as-a-judge is model-dependent: the systematic gap between Llama 3.1 and GPT-4o, together with the case-level inversions discussed above, indicates that the semantic verdict depends substantially on the evaluator model rather than on the summary itself. This suggests that single-LLM evaluation, regardless of the model chosen, provides an unstable signal of summary quality. Third, holistic dimensions do not survive the constraints imposed on semantic metrics by long legal documents. The judgments in our corpus average several thousand tokens, well beyond the input limits of the encoders typically used by metrics such as BERTScore (Zhang et al., 2020) and BARTScore (Yuan et al., 2021); the chunking strategies adopted to address this (see Zanatti, Ribeiro, Pinto, 2025) fragment precisely the global structure that dimensions like Coherence and overall Adequacy depend on. As a result, semantic metrics tend to underperform lexical ones on holistic dimensions, an effect also observed empirically in Zanatti, Ribeiro, Pinto (2025). Fourth, the human reference itself is not a unique target. The judges' free-form comments revealed disagreement about what a good summary should contain — most visibly in the discussion of article numbers, sentence connectors, and the inclusion of factual paragraphs — confirming that summary writing in the STJ context is partly subjective. Any metric that treats the human-written summary as a single ground truth inherits this subjectivity and penalizes legitimate stylistic variation.

Taken together, the observations above point away from the search for a single "best" general-purpose metric and toward an evaluation paradigm tailored to what a legal summary is actually expected to do. As noted in Section 4, Portuguese Supreme Court summaries exhibit no standard form — ranging from a single sentence to structured numbered lists. However, regardless of their form, summaries share a common functional requirement: they must contain sufficient argumentative content for the case to be intelligible, capturing the essential reasoning that connects the legal problem at stake to the decision reached. The Ac. 3421-16.3T8FNC.L1.S1 case discussed earlier illustrates why this matters: the summary attained ROUGE-1 = 0.98 and was lexically near-verbatim to the human reference, yet was criticized by the judges on grounds that no automatic metric captured. This suggests that evaluation methods for legal summarization must be grounded in the argumentative and functional properties of the domain, rather than adapted from general-purpose frameworks. A question-answering-based approach (Xu & Ashley, 2023) offers a promising path in this direction. This formulation addresses, in a single mechanism, several of the disagreements documented earlier in this section. It bypasses the long-document constraints that penalize BERTScore and BARTScore, since each comparison is performed over short, structured answer spans rather than over full judgments. It reduces the impact of stylistic variation that inflates ROUGE-1 in the legal domain, since lexical match between summary and reference is no longer the unit of comparison. And it provides a more interpretable signal than a single LLM score, since the protocol localizes failure to specific argumentative components — a summary may correctly capture the reasoning but miss the problem framing, as in Ac. 3421-16.3T8FNC.L1.S1, and the metric will say so. We plan to explore this direction in future work, building on the document structure identified by the Segmenting Judgment Model proposed in this paper and drawing on question-answering–based evaluation approaches for legal summarization (Xu & Ashley, 2023) as a foundation.

8 Limitations

While our work demonstrates the effectiveness of domain-adapted embeddings and structure-aware segmentation for extractive summarization of Portuguese Supreme Court judgments, several limitations must be acknowledged:

Dataset Size and Scope: Our summarization experiments are conducted on a corpus of 700 STJ judgments, which, while sufficient to demonstrate the effectiveness of our approach, represents a relatively small dataset by NLP standards. The limited size may affect the robustness of our conclusions and restricts the training of more data-hungry models. Future work should seek to expand the corpus, both in size and in coverage of different legal procedure categories.

Semi-Automatic Construction of Segmentation Annotations: The segmentation dataset was constructed semi-automatically, however to increase the diversity of document structures we generated additional training instances by systematically removing optional sections from existing judgments, producing variants with different structural configurations. While care was taken to ensure that original judgments and their derived variants were strictly separated across training, development, and test splits to prevent data leakage, the artificial nature of these augmented instances remains a limitation. The structural similarity between a judgment and its derived variants means that the model may not be exposed to the full range of structural variation found in real documents. Future work should address this by collecting and annotating a larger set of genuinely diverse real judgments.

Jurisdiction and Language Specificity: Our approach is specifically tailored to European Portuguese legal judgments from the STJ in terms of its trained components — the segmentation model and domain-adapted embeddings are coupled to the structural and linguistic conventions of this specific court and would require retraining to be applied to other jurisdictions or languages. However, the underlying pipeline — combining structure-aware document segmentation with domain-adapted semantic representations for extractive summarization — is designed to be generalizable. The same paradigm of identifying argumentative structure within legal documents to guide summarization has been applied to other legal corpora and jurisdictions (Xu et al., 2021), suggesting that the approach transfers across legal systems even when the trained components do not. We therefore see the jurisdiction specificity of the current system as a resource constraint rather than an architectural limitation.

LLM-Based Evaluation: Our use of Llama 3.1 and GPT-4o as evaluators, while providing a complementary perspective to ROUGE, carries known limitations. We did not perform systematic prompt engineering, and our evaluation uses a single prompt formulation per model. LLM-based evaluation is known to be sensitive to prompt design, model-specific biases, and hallucinations. The limited agreement between the two LLMs further highlights the instability of this evaluation approach. A more systematic study, calibrating LLM judgments against human evaluations in the legal domain, remains an important direction for future work.

9 Conclusion and future work

In this article, we introduce our research conducted within the framework of the *IRIS* project, a collaborative initiative involving the Portuguese Supreme Court of Justice and INESC-ID, emphasizing the importance of mutlidisciplinarity when addressing this kind of challenges. The primary contribution of this work was the creation of

domain-specific embeddings that advance the semantic modeling of Portuguese legal texts. The evaluation carried out by judges of the supreme court, while limited in scale, provides an empirical grounding that demonstrates the practical relevance of the approach. This combination of methodological innovation and domain-based validation underlines the study's empirical character and its contribution to bridging computational linguistics and judicial practice. Our efforts also encompassed the creation of a dataset from scratch and the training of the Segmenting Judgment Model to partition Portuguese Supreme Court judgments into sections, thereby enhancing our understanding of the structural aspects of these judgments. The domain-specific embeddings and the Segmenting Judgment model, combined with a simple extractive summarization technique, produced promising results that matched or surpassed those reported in related studies, while acknowledging differences in datasets and scope. We demonstrate that, for the generation of superior summaries, the summarizer benefits from context derived from the most critical section of the judgments while exclusively returning paragraphs from the pivotal section for summary production. Given that, as we have demonstrated, judges construct their summaries by selecting paragraphs from the judgments, our results indicate ample room for improvement.

In our future work, we intend to enhance our Segmenting Judgment Model by further subdividing the Grounds Section into ''facts'' and ''law''. We anticipate that by exclusively returning paragraphs from the ''law'' sub-section, we can potentially enhance our results. We also aspire to train a model that aligns more closely with the current state of the art, with the aim of maximizing the ROUGE score — a meaningful proxy given the highly extractive nature of STJ summaries. We also acknowledge the limitations of our LLM-based evaluation, including sensitivity to prompt formulation and model-specific behaviors such as hallucinations. As future work, we intend to conduct a more systematic study of LLM-based evaluation, examining its sensitivity to prompt design, model choice, and hallucinations, and calibrating LLM judgments against human evaluations in the legal domain.

A further direction concerns the evaluation methodology itself. As discussed in Section 7.4, the three evaluation signals used in this work — ROUGE, LLM-based scoring, and expert feedback — agree in broad terms but diverge at the individual summary level, and recent work on Portuguese legal summarization evaluation (Zanatti, Ribeiro, Pinto, 2025) confirms that current metrics, including those based on legal-domain embeddings, only partially capture expert judgment. Building on this evidence, we plan to develop the argument-aware, question-answering–based evaluation protocol outlined in Section 7.4, in which the Segmenting Judgment Model is repurposed to derive structure-grounded questions from a judgment against which the functional adequacy of a summary can be assessed.

CRediT authorship contribution statement

Martim Zanatti: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing – original draft, Visualization. **Rui Melo:** Conceptualization, Methodology, Software, Resources, Writing – review & editing. **Miguel Won:** Software, Resources. **Ricardo Ribeiro:** Conceptualization, Validation, Supervision, Resources, Writing – review & editing. **H. Sofia Pinto:** Supervision, Writing – review & editing. **José Borbinha:** Project administration, Supervision, Funding acquisition, Writing – review & editing. **João Dias:** Supervision, Resources, Writing – review & editing. **Pedro A. Santos:** Supervision, Resources, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

Work supported by national funds through Fundação para a Ciência e a Tecnologia, I.P. (FCT), Portugal under projects UID/50021/2025 (DOI: https://doi.org/10.54499/UID/50021/2025) and UID/PRR/50021/2025 (DOI: https://doi.org/10.54499/UID/PRR/50021/2025) and project AInT - Artificial Intelligence in Tribunal with the reference 2024.07523.IACDC (DOI: https://doi.org/10.54499/2024.07523.IACDC). This research is also part of the IRIS project with reference PR07005. This project is a collaborative involving the Portuguese Supreme Court of Justice and INESC-ID.

We would like to express our gratitude to Dr.[a] Gabriela Cunha Rodrigues and Dr.[a] Clarisse Gonçalves for their invaluable assistance in evaluating our system for the preparation of this research paper. Their expertise has been instrumental in enhancing our understanding of the Portuguese legal domain, contributing significantly to the refinement of our work.

We would also like to thank NeuralShift[31] for their valuable support in accessing the LLM-based evaluation of the STJ summaries.

Data availability

I have shared the link of the data in the manuscript.

References

Akter, M., Çano, E., Weber, E., Dobler, D., & Habernal, I. (2025). A comprehensive survey on legal summarization: Challenges and future directions. *ACM Computing Surveys*, *58*, http://dx.doi.org/10.1145/3776586.

Albayati, M., Karaoğlan, K., & Findik, O. (2025). Towards efficient multi-legal document summarization: An ensemble approach for turkish law. *Engineering Science and Technology an International Journal*, *70*.

Basyal, L., & Sanghvi, M. (2023). Text summarization using large language models: A comparative study of MPT-7b-instruct, falcon-7b-instruct, and openai chat-GPT models. CoRR abs/2310.10449, arXiv:2310.10449, URL https://doi.org/10.48550/arXiv.2310.10449.

Bird, S., Loper, E., & Klein, E. (2009). *Natural language processing with Python*. O'Reilly Media Inc.

Bowman, S. R., Angeli, G., Potts, C., & Manning, C. D. (2015). A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 conference on empirical methods in natural language processing* (pp. 632–642). Lisbon, Portugal: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/D15-1075, URL https://aclanthology.org/D15-1075.

Canaverde, B., Pires, T. P., Ribeiro, L. M., & Martins, A. F. T. (2025). LegalBench.PT: A benchmark for Portuguese law. http://dx.doi.org/10.48550/ARXIV.2502.16357, CoRR abs/2502.16357, arXiv:2502.16357.

Colombo, P., Pires, T. P., Boudiaf, M., Culver, D., Melo, R., Corro, C., Martins, A. F. T., Esposito, F., Raposo, V. L., Morgado, S., & Desa, M. (2024). SaulLM-7B: A pioneering large language model for law. http://dx.doi.org/10.48550/ARXIV.2403.03883, CoRR abs/2403.03883, arXiv:2403.03883.

Colombo, P., Pires, T. P., Boudiaf, M., Melo, R., Hautreux, G., Malaboeuf, E., Charpentier, J., Culver, D., & Desa, M. (2024). SaulLM-54B & SaulLM-141B: Scaling up domain adaptation for the legal domain. In A. Globersons, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. M. Tomczak, & C. Zhang (Eds.), *Advances in neural information processing systems 38: annual conference on neural information processing systems 2024, neurIPS 2024, vancouver, BC, Canada, December 10 - 15, 2024*. URL http://papers.nips.cc/paper_files/paper/2024/hash/ea3f85a33f9ba072058e3df233cf6cca-Abstract-Conference.html.

de Vargas Feijó, D., & Moreira, V. P. (2023). Improving abstractive summarization of legal rulings through textual entailment. *Artificial Intelligence and Law*, *31*, 91–113, URL https://doi.org/10.1007/s10506-021-09305-4.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)* (pp. 4171–4186). Minneapolis, Minnesota: Association for Computational Linguistics, URL https://aclanthology.org/N19-1423.

[31] https://www.affine.law/pt.

Endres-Niggemeyer, B., & Wansorra, E. (2004). Making cognitive summarization agents work in a real-world domain. In *Proceedings of the 1st international workshop on natural language understanding and cognitive science* (pp. 86–96). INSTICC Press.

Erkan, G., & Radev, D. R. (2004). LexRank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research*, *22*, 457–479.

Fernandes, L. C., dos Santos Ribeiro, L., de Castro, M. V. B., da Silva Pacheco, L. A., & de Oliveira Sandes, E. F. (2025). JurisTCU: A Brazilian Portuguese information retrieval dataset with query relevance judgments. http://dx.doi.org/10.48550/ARXIV.2503.08379, CoRR abs/2503.08379, arXiv:2503.08379.

Fonseca, E., Santos, L., Criscuolo, M., & Aluisio, S. (2016). ASSIN: Avaliacao de similaridade semantica e inferencia textual. In *Computational processing of the portuguese language-12th international conference, tomar, Portugal* (pp. 13–15).

Glaser, I., Moser, S., & Matthes, F. (2021). Summarization of german court rulings. In *Proceedings of the natural legal language processing workshop 2021* (pp. 180–189).

Huang, Y., Sun, L., Han, C., & Guo, J. (2023). A high-precision two-stage legal judgment summarization. *Mathematics*, *11*, 1–16, URL https://ideas.repec.org/a/gam/jmathe/v11y2023i6p1320-d1092107.html.

Huang, Z., Xu, W., & Yu, K. (2015). Bidirectional LSTM-CRF models for sequence tagging. CoRR abs/1508.01991, arXiv:1508.01991, URL http://arxiv.org/abs/1508.01991.

Jain, D., Borah, M. D., & Biswas, A. (2023). Bayesian optimization based score fusion of linguistic approaches for improving legal document summarization. *Knowledge-Based Systems*, *264*, Article 110336, URL https://doi.org/10.1016/j.knosys.2023.110336.

Jain, D., Borah, M. D., & Biswas, A. (2024). A sentence is known by the company it keeps: Improving legal document summarization using deep clustering. *Artificial Intelligence and Law*, *32*, 165–200, URL https://doi.org/10.1007/s10506-023-09345-y.

Li, T., Li, Z., & Zhang, Y. (2024). Improving faithfulness of large language models in summarization via sliding generation and self-consistency. In N. Calzolari, M. Kan, V. Hoste, A. Lenci, S. Sakti, & N. Xue (Eds.), *Proceedings of the 2024 joint international conference on computational linguistics, language resources and evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy* (pp. 8804–8817). ELRA and ICCL, URL https://aclanthology.org/2024.lrec-main.771.

Lin, C. Y. (2004). ROUGE: A package for automatic evaluation of summaries. In *Text summarization branches out* (pp. 74–81). Barcelona, Spain: Association for Computational Linguistics, URL https://aclanthology.org/W04-1013.

Liu, Y., & Lapata, M. (2019). Text summarization with pretrained encoders. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing* EMNLP-IJCNLP, (pp. 3730–3740). Hong Kong, China: Association for Computational Linguistics, URL https://aclanthology.org/D19-1387.

Luhn, H. P. (1958). The automatic creation of literature abstracts. *IBM Journal of Research and Development*, *2*, 159–165, URL https://doi.org/10.1147/rd.22.0159.

May, P. (2021). Machine translated multilingual STS benchmark dataset. https://github.com/PhilipMay/stsb-multi-mt, URL https://github.com/PhilipMay/stsb-multi-mt.

Melo, R., Santos, P. A., & Dias, J. (2023). A semantic search system for the supremo tribunal de justiça. In N. Moniz, Z. Vale, J. Cascalho, C. Silva, & R. Sebastião (Eds.), *Progress in artificial intelligence - 22nd EPIA conference on artificial intelligence, EPIA 2023, Faial Island, Azores, September 5-8, 2023, Proceedings, Part II* (pp. 142–154). Springer, URL https://doi.org/10.1007/978-3-031-49011-8_12.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. In Y. Bengio, & Y. LeCun (Eds.), *1st international conference on learning representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop track proceedings*. URL http://arxiv.org/abs/1301.3781.

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing* EMNLP, (pp. 1532–1543). Doha, Qatar: Association for Computational Linguistics, URL https://aclanthology.org/D14-1162.

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., & Manning, C. D. (2020). Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th annual meeting of the association for computational linguistics: System demonstrations*. URL https://nlp.stanford.edu/pubs/qi2020stanza.pdf.

Real, L., Fonseca, E., & Oliveira, H. G. (2020). The ASSIN 2 shared task: A quick overview. In *International conference on computational processing of the portuguese language* (pp. 406–412). Springer.

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of the 2019 conference on empirical methods in natural language processing*. Association for Computational Linguistics, https://arxiv.org/abs/1908.10084.

Ribeiro, R., & Martins de Matos, D. (2008). Mixed-source multi-document speech-to-text summarization. In *Coling 2008: proceedings of the workshop multi-source multilingual information extraction and summarization* (pp. 33–40). Manchester, UK: Coling 2008 Organizing Committee, URL https://aclanthology.org/W08-1406.

Souza, F., Nogueira, R., & Lotufo, R. (2020). BERTimbau: Pretrained BERT models for Brazilian Portuguese. In R. Cerri, & R. C. Prati (Eds.), *Intelligent systems* (pp. 403–417). Cham: Springer International Publishing.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. In I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, & R. Garnett (Eds.), *Advances in neural information processing systems 30: Annual conference on neural information processing systems 2017, December 4-9, 2017, Long Beach, CA, USA* (pp. 5998–6008). URL https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Wang, K., Thakur, N., Reimers, N., & Gurevych, I. (2022). GPL: Generative pseudo labeling for unsupervised domain adaptation of dense retrieval. In M. Carpuat, M. de Marneffe, & I. V. M. Ruíz (Eds.), *Proceedings of the 2022 conference of the North American chapter of the association for computational linguistics: Human language technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022* (pp. 2345–2360). Association for Computational Linguistics, http://dx.doi.org/10.18653/V1/2022.NAACL-MAIN.168.

Xu, H., & Ashley, K. D. (2023). A question-answering approach to evaluating legal summaries. In G. Sileno, J. Spanakis, & G. van Dijck (Eds.), *Legal knowledge and information systems - JURIX 2023: The thirty-sixth annual conference, maastricht, The Netherlands, 18-20 December 2023* (pp. 293–298). IOS Press, http://dx.doi.org/10.3233/FAIA230977.

Xu, H., Savelka, J., & Ashley, K. D. (2021). Toward summarizing case decisions via extracting argument issues, reasons, and conclusions. In J. Maranhão, & A. Z. Wyner (Eds.), *ICAIL '21: eighteenth international conference for artificial intelligence and law, São Paulo Brazil, June 21 - 25, 2021* (pp. 250–254). ACM, http://dx.doi.org/10.1145/3462757.3466098.

Yuan, W., Neubig, G., & Liu, P. (2021). Bartscore: Evaluating generated text as text generation. In M. Ranzato, A. Beygelzimer, Y. N. Dauphin, P. Liang, & J. W. Vaughan (Eds.), *Advances in neural information processing systems 34: annual conference on neural information processing systems 2021, NeurIPS 2021, December 6-14, 2021, Virtual* (pp. 27263–27277). URL https://proceedings.neurips.cc/paper/2021/hash/e4d2b6e6fdeca3e60e0f1a62fee3d9dd-Abstract.html.

Zanatti, M., Ribeiro, R., & Pinto, H. S. (2024). Segmentation model for judgments of the Portuguese supreme court of justice. In M. F. Santos, J. Machado, P. Novais, P. Cortez, & P. M. Moreira (Eds.), *Progress in artificial intelligence - 23rd EPIA conference on artificial intelligence, EPIA 2024, Viana Do Castelo, Portugal, September 3-6, 2024, Proceedings, Part I* (pp. 245–257). Springer, URL https://doi.org/10.1007/978-3-031-73497-7_20.

Zanatti, M., Ribeiro, R., & Pinto, H. S. (2025). Exploring metric correlations for legal text summarization evaluation. In J. Maranhão (Ed.), *Proceedings of the twentieth international conference on artificial intelligence and law, ICAIL 2025, Chicago, IL, USA, June 16-20, 2025* (pp. 389–393). ACM, http://dx.doi.org/10.1145/3769126.3769206.

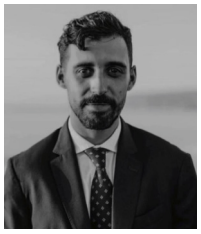
Zanatti, M., Ribeiro, R., Pinto, H. S., & Borbinha, J. (2025). Atribuição de descritores a acórdãos do supremo tribunal de justiça Português com base em representações locais esparsas. *Linguamática*, *17*, 35–51. http://dx.doi.org/10.21814/LM.17.1.481.

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2020). Bertscore: Evaluating text generation with BERT. In *8th international conference on learning representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, URL https://openreview,net/forum?id=SkeHuCVFDr.

Zhang, T., Ladhak, F., Durmus, E., Liang, P., McKeown, K. R., & Hashimoto, T. B. (2024). Benchmarking large language models for news summarization. *Transactions of the Association for Computational Linguistics*, *12*, 39–57, URL https://doi.org/10.1162/tacl_a_00632.

Zhang, H., Yu, P. S., & Zhang, J. (2025). A systematic survey of text summarization: From statistical methods to large language models. *ACM Computing Surveys*, *57*, http://dx.doi.org/10.1145/3731445.

Zhong, L., Zhong, Z., Zhao, Z., Wang, S., Ashley, K. D., & Grabmair, M. (2019). Automatic summarization of legal decisions using iterative masking of predictive sentences. In *Proceedings of the seventeenth international conference on artificial intelligence and law, ICAIL 2019, Montreal, QC, Canada, June 17-21, 2019* (pp. 163–172). ACM, URL https://doi.org/10.1145/3322640.3326728.



Martim Zanatti is a PhD student at IST - Instituto Superior Técnico and serves as a researcher at INESC-ID. His doctoral research focuses on Natural Language Processing (NLP) with a specialization in legal documents, particularly Portuguese legal texts. His work encompasses various tasks including summarization, Extreme Multi-label Classification (XML), legal document segmentation, summarization evaluation methods, and the creation of legal embeddings for Portuguese language. He is a member of the Center for Responsible AI and was a member of the IRIS project, a collaborative initiative between INESC-ID and the Supreme Court of Justice of Portugal. His contributions to these projects aim to enhance the application of NLP techniques in the legal domain, improving the efficiency and accuracy of legal document processing. He holds a Master's degree from Instituto Superior Técnico, where his thesis centered on music, specifically the automatic generation of guitar tablatures. His interdisciplinary background combines expertise in both the technical aspects of NLP and practical applications in diverse fields such as law and music.



Rui Melo is a PhD student jointly at Carnegie Mellon University (USA) and the University of Porto (Portugal), where he researches the intersection of Machine Learning and Software Engineering. He holds an MSc from Instituto Superior Técnico and previously worked as an AI Engineer at a U.S. legal-tech startup. His earlier research focused on applying Machine Learning to the legal domain, particularly in developing encoder models tailored to Portuguese law. His current work explores AI safety in code-related tasks, namely in agentic workflows and through the use of mechanistic interpretability techniques.



Miguel Won is an AI Engineer at Axians Portugal. He holds a PhD in Particle Physics from the University of Coimbra, where he focused on computational methods for high-energy physics. Prior to joining industry, he worked as a Postdoctoral Researcher at INESC-ID, in the areas of Natural Language Processing (NLP) and Computational Social Science. His research explored the analysis of political communication through parliamentary speeches and social media dynamics, leading to several peer-reviewed publications in international journals such as the Journal of Computational Social Science and Political Communication. His current interests lie at the intersection of AI engineering, NLP, Agentic AI and data-driven social analysis.



Ricardo Ribeiro (PhD) is an Associate Professor at Iscte - Instituto Universitário de Lisboa and an integrated researcher at INESC-ID Lisboa, working on Human Language Technologies. He has participated in several European and Nationally-funded projects, being the Human Language Technologies INESC-ID Team Coordinator at RAGE (2015–2019) European-funded project and a principal investigator of a Ministry of National Defense-funded project on information extraction from text. He has been member of the program committee of several main scientific events on Artificial Intelligence and Natural Language Processing, such as IJCAI, ECAI, ICASSP, LREC, and Interspeech. He was one of the editors of a book on the computational processing of Portuguese language. His current research interests focus on high-level information extraction from unrestricted text, speech, or music, and improving machine-learning techniques using domain-related information.



H. Sofia Pinto is an Assistant Professor at the Department of Computer Science and Engineering of Instituto Superior Técnico, University of Lisbon. She holds a PhD in Computer Science and Engineering, in the area of Artificial Intelligence. She is a Researcher at INESC-ID (PI). She has worked in knowledge representation & reasoning and semantic web: ontologies and ontology engineering, inheritance systems, and truth maintenance systems. Her research interests include ontology engineering, semantics and social media for organizational learning, social network influence analysis (complex networks), computer vision. In the last years she has been researching in computational creativity, support for and identification of mental illnesses (anxiety, depression, psychosis, bipolar), creative and personalized fake news explanations. She is also researching in the use of AI in several NLP problems namely summarization, segmentation and categorization of supreme court judgments. She has published more than 60 refereed scientific papers on those areas and received two best paper awards. She was the (co)-Program Chair of EKAW 2010, EPIA 2011, I-SEMANTICS 2012, STAIRS 2016. She served in several Chair or Committee roles for several ISWC, ESWC, The Web Conf, KRR, ECAI, ICCC, IJCAI, ER.



José Borbinha is Full Professor in the Department of Computer Science and Engineering at Instituto Superior Técnico, University of Lisbon, and a Senior Researcher at INESC-ID. His main research interests lie in the engineering of information systems, particularly in the requirements, analysis, and design of information systems. His work focuses on topics such as information management, digital libraries, archives, records management, and digital preservation. From 1998 to 2005, he served as Chief Information Officer of the National Library of Portugal, where he played a key role in the early initiatives that led to the creation of Europeana.eu, the European discovery service integrating resources from libraries, archives, museums, and publishers. He was also a founding member of the IEEE Technical Committee on Digital Libraries, serving as its elected chair from 2008 to 2010.



João Dias is an Assistant Professor at Faculty of Science and Technology, from University of Algarve. He is currently the Coordinator and Integrated Researcher at CISCA (Centro de Investigação em Sistemas Ciberfísicos do Algarve), and colaborator at INESC-ID. He has over 70 peer-reviewed publications and has participated in 9 research projects, 7 of which were international research projects. He is also currently the Director of the Master in Informatics Engineering at University of Algarve. His research focus on the design and development of Intelligent Interactive Systems, including Natural Language Processing, Intelligent Virtual Agents and Serious Games.



Pedro A. Santos is Associate professor at the Mathematics Department, and the Computer Science and Engineering Department, of Instituto Superior Técnico – University of Lisbon. He is also a Senior Researcher at INESC-ID. He has a doctoral degree in Mathematics. His current research interests and contributions go from Mathematical foundations of Machine Learning and Artificial Intelligence, to Game Design and Game Studies. He is a co-author of books on Non-Commutative Gelfand Theories (2011), on Game Design and Development (2014), on Operator Algebras (2019), and is author and co-author of more than 70 peer-reviewed papers published in scientific journals or in conference proceedings.