SARA - A Generative AI for Legal Process Summarization Based on Chain of Density Prompt Engineering

Francisco das Chagas Jucá Bomfim[1[0000−0001−6160−7832]], Joao Araujo Monteiro Neto[1[0000−0002−0690−2449]], Gilson Bezerra Filho[1[0009−0004−6868−7758]], Vasco Furtado[1[0000−0001−8721−4308]], and Vládia Pinheiro[1[0000−0002−9851−8304]]

Universidade de Fortaleza, Fortaleza, CE, Brasil
fjuca@unifor.br, joaoneto@unifor.br, gilsonbezerram@edu.unifor.br, vasco@unifor.br, vladiacelia@unifor.br
http://www.unifor.br

Abstract. Generative AI, particularly Large Language Models (LLMs), holds significant promise for enhancing judicial tasks, especially in automating the generation of legal cases during the sentencing phase. This paper introduces SARA (System for Analysis and summaRization of legal Actions), an innovative method for abstractive and multi-document summarization of legal proceedings. SARA employs GPT-4o, trained exclusively through in-context learning, utilizing Chain of Density (CoD) and CO-STAR prompt engineering techniques. These methods, adapted from judicial procedural knowledge, significantly improve the quality of generated summaries. Traditional evaluation metrics reveal effective training strategies but highlight their limitations in assessing summary quality. Therefore, we propose a qualitative evaluation methodology based on expert-generated questionnaires, focusing on essential content inclusion and proper report structuring. Inspired by this methodology, we trained another GPT model for large-scale summary evaluation. Evaluations of summaries of fifteen first-degree court cases from the Court of Justice of the State of Ceará show a significant advantage of in-context learning with CoD, emphasizing the role of domain knowledge and report style.

Keywords: Text Summarization · Legal Documents · Generative AI · Prompt Engineering.

1 Introduction

Generative Artificial Intelligence (GAI), particularly Large Language Models (LLMs), exhibits remarkable language generation capabilities shows great potential for various judicial tasks. Recently, the Federal Superior Court issued public call 01/2023 to develop AIs tailored for these purposes [17], aligning with the goals set by the CNJ (National Council Of Justice) to enhance procedural efficiency and adhere to global judicial innovation practices [12]. A key objective is the automatic generation of the process report, crucial for ensuring judges are well-informed about essential case issues, thereby justifying their final decisions adequately.

A process report must accurately narrate the facts and allegations presented, documenting the parties involved, their arguments, summaries of requests and defenses, and major events during the process. Producing such a report is a complex, time-consuming task that requires domain knowledge and synthesis ability, making it well-suited for abstractive summarization. Unlike extractive summarization, which merely selects phrases from the text, abstractive summarization generates a coherent synthesis that captures the essence of the case.

Research on using LLMs for abstractive summarization of legal documents has progressed, but it faces challenges, such as handling the length and volume of legal texts and evaluating summary quality, which often relies on syntactic metrics that inadequately capture semantic context [11]. Notable State-of-the-Art (SOTA) models include SimCLS [15] for English, which employs Contrastive Learning (CL), LegalSumm for Portuguese legal documents [13], and CLSJUR.BR [14], which combines CL with free-reference evaluation techniques and a language model for legal documents in Brazilian Portuguese - LegalBERT.PT [19]. These models, however, have limitations related to the original document size and the nature of the generated summaries, typically producing only syllabuses. Traditional metrics like ROUGE and BLEU, though useful for comparison, fall short of accurately reflecting how well a summary captures the essence of the document.

In this paper, we introduce SARA (System for Analysis and summaRization of legal Actions), a novel method for abstractive and multi-document summarization of legal proceedings aimed at generating judicial reports. SARA's architecture is based on an LLM (GPT-4o [1]) trained exclusively through in-context learning. This training employs two prompt engineering techniques: Chain of Density (CoD) [2] for summarization and CO-STAR [3] for structuring prompt content. We demonstrate that these techniques, adapted from procedural knowledge provided by judges and legal academics, significantly enhance the informative quality of the generated summaries.

Results using traditional model quality assessment metrics indicate which training strategies are most effective, but they also highlight the limitations of these metrics in evaluating summary quality. Consequently, we propose a qualitative evaluation methodology based on expert-generated questionnaires. This methodology focuses on essential content inclusion and the appropriate structuring of process reports.

Beyond guiding prompt engineering and enabling human-quality analysis of summaries, this methodology inspired the training of ChatGPT[6] to evaluate generated summaries, facilitating large-scale assessments. Evaluations of summaries from fifteen first-degree judicial cases in a large Brazilian State Court, conducted by both machines and humans, reveal a significant advantage of summaries produced using in-context learning with CoD, enhanced by domain knowledge and report style familiarity.

2 Related Work

Automatic summarization can be classified into two types: abstractive and extractive. Abstractive summarization generates new text that captures the original content's essence, often rephrasing it concisely. Extractive summarization selects direct excerpts from the original text. Summarization can also be categorized as single-document or multi-document, with the latter integrating information from multiple sources into a cohesive summary. For first-degree judicial processes, the most suitable automatic summarization is abstractive and multi-document.

Several works before the GPT era proposed various approaches to legal text summarization. For example, [15] employed Contrastive Learning for abstractive summarization, while [7] used Deep Reinforcement Learning, combining extractive and abstractive techniques. "DELSumm: Incorporating Domain Knowledge for Extractive Summarization of Legal Case Documents" [10] utilized domain-specific knowledge for extractive summarization, achieving high ROUGE metrics (ROUGE-1: 51.2) but requiring manual annotations. A similar approach was attempted by [9], who enhanced summarization by incorporating markers like <Issue> and <Argument>, leading to improved ROUGE scores (ROUGE-1: 50.73). However, manual tagging presents a scalability limitation.

For Brazilian Portuguese, LegalSumm [13] employs Transformer models [16] and Contrastive Learning to summarize judicial decisions from Brazil's Supreme Federal Court, generating concise summaries. Using the Ruling.BR [13] corpus for training and testing, results on ROUGE metrics demonstrated that LegalSumm improved the quality of generated summaries compared to traditional methods. In [14], the authors proposed the CLSJUR.BR model, an evolution of LegalSumm, which uses the CL approach along with a new technique for generating and evaluating candidate summaries. CLSJUR.BR is based on the MBart neural model [18] and the LegalBERT.PT language model [19]. The results, using the same LegalSumm test set, showed an improvement in the generation of syllabuses with ROUGE-L from 0.35 (LegalSumm) to 0.4789 (CLSJUR.BR). The limitation of CLSJUR.BR for generating court reports is the limitation for mono documents and the size of the input text (1024 tokens).

In summary, while research on abstractive and multi-document automatic legal document summarization (ALDS) has made advances with SimCLS, LegalSumm, and CLSJUR.BR models, significant challenges remain due to the complexity and volume of legal texts. Recent work consistently highlights the need to explore the potential of Generative AI (GAI) and Large Language Models (LLMs) to improve the efficiency and accuracy of judicial report generation, contributing to a more agile and well-founded justice system.

3 SARA - Generating the Report Process

In this section, we introduce the architecture of SARA, a GAI designed for the analysis and summarization of legal cases, and the prompt engineering process that led to the final prompt for multi-document abstractive summarization of legal cases. SARA specifically generates the section known as the Report in the Brazilian judicial system, whose main objective is to summarize the key factual and legal elements of the procedural relationship, including the names of the parties, the identification of the case with a summary of the request and objection, and a documentation of the main events that occurred during the process.

A critical aspect of court reporting is the ability to summarize from multiple documents in a cohesive and accurate manner. Therefore, the core of SARA is a multi-document summary prompt capable of integrating information from diverse sources, including initial pleadings, responses, replies, terms of hearing, closing briefs, and other documents within the judicial process that are essential for understanding the legal demand and reaching a decision in the case. This efficient summarization process is essential for procedural speed, allowing an integrated and holistic view of cases.

3.1 SARA Architecture

Figura 1 presents SARA's basic architecture [`https://l1nk.dev/HPiPv`] with the following modules:

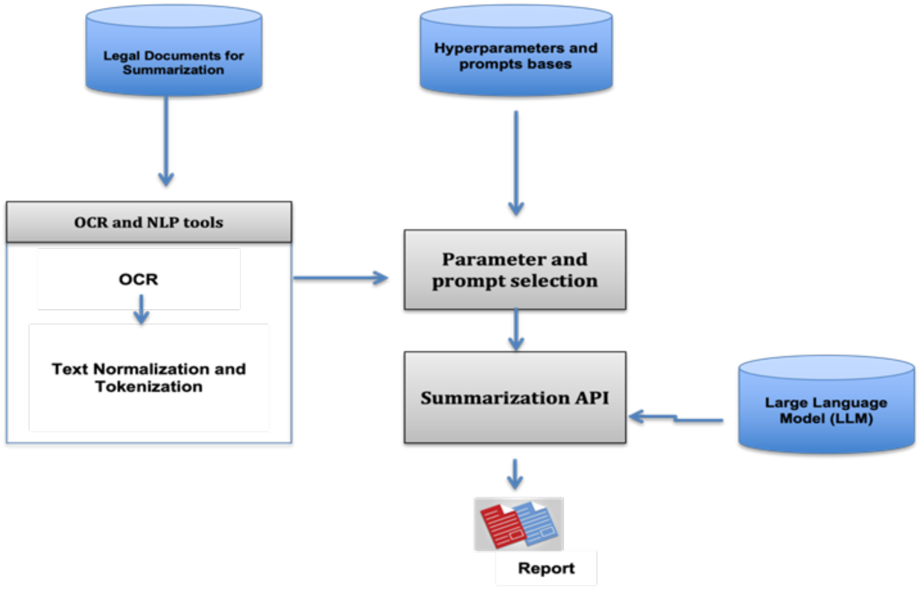
Legal Documents for Summarization
Hyperparameters and prompts bases
OCR and NLP tools
OCR
Text Normalization and Tokenization
Parameter and prompt selection
Summarization API
Large Language Model (LLM)
Report

Fig. 1. SARA's General Architecture

- **OCR and NLP Tools -** responsável por realizar o reconhecimento óptico de caracteres (OCR) e verificações como validade do documento em relação ao tamanho, linguagem válida e normalização dos múltiplos documentos de entrada;

- **Parameter and Prompt Selection** - responsável por selecionar o prompt a ser usado na sumarização e configurá-lo com base nos seguintes hiperparâmetros: lista de documentos a serem resumidos `<DOCS LIST>`; modelo padrão do relatório judicial a ser gerado `<REPORT TEMPLATE>`; tamanho inicial do relatório `<T>`; número `<N>` de etapas no processo de densificação (estratégia CoD); número `<M>` de entidades ausentes; lista de informações obrigatórias que devem ser incluídas no resumo `<REQUIRED-INFO>`;
- **Summarization API** - rotina que envia a instrução de sumarização (prompt) para o LLM junto com os documentos dos casos a serem resumidos.

3.2 Versions of SARA's Prompts

Prompt engineering is an essential skill, focusing on creating clear and precise instructions to guide LLMs in generating the desired outcomes [5]. To instruct SARA in generating high-quality judicial process reports, an incremental approach to prompt engineering was used, focusing on enhancing the following qualitative aspects of judicial reports: form, clarity, precision, and density. In total, four versions of prompts were developed following the CO-STAR framework [3] and the CoD strategy [2].

The CO-STAR framework offers an organized and consistent structure for prompts, dividing them into layers of form and content to facilitate the understanding and analysis of information. The form layers include CONTEXT, which provides an overview of the scenario; OBJECTIVE, which defines the task's purpose; STYLE, which specifies the type of language; TONE, which indicates the desired attitude or emotion; AUDIENCE, which identifies the recipients; and RESPONSE, which presents the expected final format. he content layers include specific details about the documents and the legal scenario, explanations of the mandatory elements in the summary, and guidelines on appropriate language and register, ensuring that the legal summaries are accurate, coherent, and well-organized to meet the needs of judges and other legal professionals involved in the case.

On the other hand, the CoD (Chain of Density) prompt engineering strategy ensures that every word in the text significantly contributes to the overall content, removing redundancies and superfluous information. This technique has proven effective in increasing the informational density of summaries, improving clarity and conciseness. The generic densification process is detailed below:

> Step 1: Identification of Missing Entities Identify from one to <M> missing informative entities that are absent in the previously generated summary.
> Repeat step 2 for <N> times.
> Step 2: Write a new, denser summary of similar length that covers all entities and details from the previous summary, as well as the missing entities. The final summary must contain only the essential text of the story, without any instructions, steps, or guidelines. A missing entity is: relevant to the main story; specific - descriptive but concise (6 words or

> less); new - not in the previous summary; faithful - present in the original documents; located anywhere in the original documents. Avoid specific terms that are not generalizable; use standardized language according to court thesauri;

This section outlines the evolution of the SARA prompts, with each version improving accuracy, clarity, and density in generated summaries. Reports from each prompt version were qualitatively evaluated by three judges from the Court of Justice of Ceará. The evaluation revealed that the initial use of CoD and CO-STAR techniques failed to meet judicial report requirements, such as maintaining chronological order and adhering to legal guidelines. To address this, a template was introduced in the prompt, ensuring facts were presented in the correct order and format, incorporating domain knowledge and legal style.

The *Prompt-CoD-baseline* represents the starting point, using the CoD strategy proposed in [2] to generate dense summaries, but it lacked the necessary nuances and precision to capture the complexity of legal texts [9].

The next version, called *Prompt-CoD-Summ-zeroShot*, applied zero-shot learning techniques, allowing the model to generalize its knowledge from other domains to the specific task of legal summarization without additional training on the legal corpus. The hyperparameters for this version were defined by empirical experiments from the STF 01/2023 public call [17]. The CONTEXT and OBJECTIVE sections specify which documents will be processed by SARA and what the AI should do, outlining the mandatory information to include in the summary, such as the nature of the action, urgent relief, involved parties, legal arguments, relevant citations, and requests made by the parties.

The GUIDELINES section provides detailed instructions for executing the summary based on the Chain of Density (CoD) strategy. The essential segments of this section are:

> 1) They provide guidance on maintaining the most relevant information, ensuring that it is presented clearly and comprehensibly for anyone.
> 2) They define the summary length (T=1000 words), with an emphasis on clarity and informational density.
> 3) To ensure informational density, it is instructed that the summary be rewritten N=5 times to improve the flow and make room for additional entities. Less informative sentences should be merged, compressed, or removed to make the text more compact and meaningful. The following excerpt from the prompt, (GUIDELINES section), is responsible for this densification task:
> "Follow these steps:
>
> *Sample Heading (Fourth Level)* Step 1: Identify 1 to 5 missing informative entities (delimited by ";") that are absent in the previously generated summary.
> Repeat Step 2 five times.
> Step 2: Write a new, denser summary of similar length that covers all entities and details from the previous summary, as well as the missing

> entities. The final summary must contain only the essential text of the story, without any instructions, steps, or guidelines."

These guidelines aim to ensure that the summaries are clear, concise, densely informative, and faithful to the original content. Continuous rewriting of the summary, according to CoD instructions, is essential to maintain the density and quality of the information, facilitating the understanding and use of the summarized documents.

The other sections, STYLE, TONE, AUDIENCE, and RESPONSE, aim to guide the language model in generating formal and professional texts with an assertive and informative tone, suitable for legal professionals, and with a logical narrative free of value judgments.

The qualitative evaluation of the reports generated by SARA using the Prompt-CoD-Summ-zeroShot version indicated that, despite their overall quality, they exhibited the following issues: (1) Form: They do not meet legal formalities, as they do not follow the chronological order of events in the process; (2) Clarity: Important information is omitted, such as the minutes of the evidentiary hearing and the Public Prosecutor's opinion; (3) Accuracy: Only one of the parties' requests is mentioned, whereas there were multiple requests in the process.

These issues are critical and can cause the nullity of the judgment. If the order of events is unclear or if there are relevant omissions, it can result in a violation of the adversarial principle. If the facts are not presented in an organized and complete manner, the decision may be considered insufficiently substantiated. Additionally, if the narration of the facts is confusing and disorganized, the defense may be unable to present arguments or adequately contest the points raised.

The Prompt-CoD-Summ-DomainKnowledge version was designed to enhance report quality by incorporating legal domain knowledge, ensuring compliance with judicial standards, emphasizing clear citations, organized information, and the use of standardized legal language. The *Prompt-CoD-Summ-DK-Styles* version further advanced this by integrating domain knowledge with specific writing styles, introducing command and content layers, and using a REPORT-TEMPLATE developed by magistrates. This version focused on accuracy and presentation, receiving the highest evaluations from magistrates for its clarity, form, and content accuracy.

Table 1 presents the final structure of SARA's prompt, indicating the hyperparameters to be instantiated.

Section	Instructions and Hyperparameters
Version	Prompt-CoD-Summ-DK-Styles
CONTEXT	You will receive a set of legal documents related to a case. These documents include `<DOCS_LIST>`.
OBJECTIVE	Summarize legal documents using the `<REPORT_TEMPLATE>`, write the report conservatively, maintaining fidelity to the information in the original documents, act with precision and clarity, ensuring that all important information is preserved, generating a report for a judicial decision. The summary must include the following mandatory information: `<REQUIRED_INFO>`. All information about the plaintiff must come before the defendant's information. If any of this information is not available in the documents, note its absence in the summary.
STYLE	The writing style should be formal and professional, using appropriate legal terminology. Summaries and the report should be clear, direct, and objective, allowing quick understanding of legal issues. In narrative form.
TONE	The tone should be assertive and informative, reflecting the seriousness and importance of the legal content. Accuracy is essential, ensuring that all information presented is precise and relevant to the case.
AUDIENCE	The target audience includes judges and other legal professionals involved in the case, including members of the legal team or other court staff. The report should meet their need for concise but comprehensive information to support legal and strategic decision-making.
RESPONSE	The final report should be in narrative form, without making value judgments or opinions about the case, strictly adhering to the summary, and under no circumstances structured in topics, maintaining logic, starting with an introduction to the case, followed by document summaries.
GUIDELINES	The summary should ensure clarity in writing, allowing people outside the legal field to understand it; be concise, focusing on the essential without omitting significant details; be faithful to the content of the original vote, ensuring complete coverage of essential elements. The first summary should be long (approximately `<T>` words), mainly covering all important entities and containing little information beyond the defined entities. Use overly detailed language and fillers (e.g., `"this text contains"`) to reach about `<T>` words. Make every word count: rewrite the previous summary to improve flow and make room for additional entities. It is essential to indicate the legislation cited in the reference text in the summary, with appropriate reference to the original text. Follow these steps: Step 1: Identify 1 to `<M>` missing informative entities (delimited by `";"`) that are absent in the previously generated summary. Repeat Step 2 for `<N>` times. Step 2: Write a new, denser summary of similar length that covers all entities and details from the previous summary, as well as the missing entities. Avoid specific terms that are not generalizable; use standardized language according to court thesauri.

Table 1. General Structure of the SARA Abstractive Summarization Prompt Prompt-CoD-Summ-DK-Styles

4 Empirical Evaluation

The experimental evaluation consisted of generating reports for 15 first-degree cases from the Court of Justice of the state of Ceará by SARA, using the following versions of prompts: V0 - Prompt-CoD-baseline; V1 - Prompt-CoD-Summ-zeroShot; V2 - Prompt-CoD-Summ-DomainKnowledge; V3 - Prompt-CoD-Summ-DK-Styles. The reports were then compared with the reference report using traditional ROUGE [20] and BLEU [21] metrics, as the available metrics are not yet sufficiently reliable for this purpose [4,6]. Finally, a qualitative evaluation was performed by the human experts and by LLM GPT-4o using an evaluation prompt designed in this study and a questionnaire formulated by expert magistrates from the Court of Justice of the state of Ceará.

4.1 Quantitative Evaluation

Table 2 presents the average results of the ROUGE (F1-Score) and BLEU metrics for the reports of the processes generated by SARA with prompt versions V0 to V3, compared to the reference report produced by the magistrates. Individual results for each report are available at `https://l1nk.dev/HPiPv`.

In summary, version V2 performed best in ROUGE-1 F1, showing better word accuracy. Version V3 excelled in ROUGE-2 F1 and ROUGE-L F1, capturing more bigrams and preserving sentence structure and fluency. While V2 had the highest BLEU score, it may have been penalized for incorrect sequences. Version V0 had the lowest scores, struggling with both word accuracy and structure. Despite a lower BLEU score, V3 demonstrated superior structural quality and coherence.

Prompt Version	ROUGE-1 F1	ROUGE-2 F1	ROUGE-L F1	BLEU Score
V0	0.230088	0.088335	0.123917	0.043988
V1	0.421370	0.173248	0.178351	0.049367
V2	0.436677	0.180113	0.180793	0.055854
V3	0.380506	0.182838	0.192885	0.039577

Table 2. Average ROUGE (F1-Score) and BLEU Metrics for the different prompt versions.

4.2 Qualitative Evaluation

The quantitative metrics (see Table 2) demonstrated weak correlations between the reports generated by SARA and those generated by the magistrates, being insufficient to decide on one prompt strategy over the other. In this work, we conducted two qualitative evaluation strategies to capture quality criteria for the generated reports.

In the first strategy, three magistrates evaluated all reports generated by the prompt versions V0 - Prompt-CoD-baseline and V1 - Prompt-CoD-Summ-zeroShot, based on the following criteria: form, accuracy, clarity, and density. All reports were rejected on all criteria, being considered incorrect and insufficient to support judicial decisions. Interestingly, version V1 achieved one of the best results in the ROUGE-1 F1 metric but was rejected by the experts. To address the quality issues, domain knowledge and style information (template) were incorporated into the subsequent versions V2-Prompt-CoD-Summ-DomainKnowledge and V3- Prompt-CoD-Summ-DK-Styles. Reports were generated by these versions and evaluated by the magistrates. Version V2 fully met the criteria of form, accuracy, and density, and partially met the clarity criterion due to the absence of some legal reasoning. Version V3 fully met all the evaluated criteria.

In a second strategy, a qualitative questionnaire consisting of 13 questions was developed (Table 3) by the magistrate experts, with the responses ranging from 1 (report does not meet) to 5 (report fully meets). In order to systematize and automate a qualitative evaluation of reports on a large scale, focusing on form, clarity, accuracy, and utility, the following prompt was submitted to ChatGPT 4o along with the questionnaire and five versions of reports for each case in the sample: the reference report and four summaries generated by SARA using the four prompt versions V0 to V3:

> "you are a judicial analyst who will receive 5 versions of judicial process summaries and should evaluate them according to a provided <questionnaire>. At the end, you should indicate the summary that best answers the questions in the evaluation questionnaire and justify your response. <summary1> <summary2> <summary3> <summary4> <summary5> <questionnaire>”

Item	Topic	Question	V0	V1	V2	V3
1	Identification of the Parties	Were the parties involved in the judicial action clearly identified?	4.200	4.900	4.900	5.000
2	Summary of the Claim	Was the claim or claims made by the plaintiff in the initial petition clearly summarized?	4.100	4.800	4.900	5.000
3	Summary of the Defense	Was the defense presented by the defendant clearly summarized?	4.000	4.700	4.800	5.000
4	Contextualization of the Facts	Were the main facts alleged by the parties clearly contextualized?	4.300	4.900	4.900	5.000
5	Legal Background of the Initial Petition	Were the legal background of the initial petition properly summarized?	3.900	4.400	4.500	5.000
6	Legal Background of the Contestation	Were the legal background of the refutation properly summarized?	4.000	4.500	4.600	5.000
7	Relevant Procedural Occurrences	Were the main occurrences in the process clearly described?	4.200	4.800	4.800	5.000
8	Produced Evidence	Was the evidence produced by both parties clearly described?	4.100	4.900	4.900	5.000
9	Cited Legal Norms	Were the legal norms cited by the parties clearly identified?	4.200	4.900	4.900	5.000
10	Legal Reasoning	Was the application of legal norms to the facts of the case accurately presented?	4.300	4.800	4.900	5.000
11	Coherence with Arguments and Evidence	Is the report consistent with the arguments and evidence presented by the parties?	4.200	4.900	4.900	5.000
12	Chronological Order	Was the report written in chronological order?	4.100	4.800	4.800	5.000
13	Preparation for Decision	Did the report adequately prepare the ground for the reasoning and dispositive part of the judgment?	4.300	4.900	4.900	5.000

Table 3. Qualitative Evaluation Questionnaire and average of answers given by ChatGPT 4.0 for the questionnaire, for each report generated by SARA's prompt versions V0 to V3.

Versions V1 and V2 were fully reproved by the human experts, even though V1 was not so bad in quantitative metrics. Among the versions V2 and V3, version V3 stood out in the human evaluation on all the criteria established by the evaluation questionnaire, as well as by the GPT-4o language model in ratings, with the language model choosing version V3 based on criteria defined in the evaluation questionnaire. Among the findings, we can highlight that, although all versions demonstrated clarity in identifying the involved parties, version V3 stood out by maintaining this clarity throughout the entire document. This aspect significantly facilitates the reader's understanding, as evidenced by the consistent high ratings in all evaluations of "Identification of the Parties" in the analyzed versions.

Version V3 provided a detailed and accurate summary of both the claim and the defense, covering all essential aspects without significant omissions. For example, while versions V1 and V2 received high evaluation for clarity, both presented omissions in important specific details. Version V3, on the other hand, ensured the complete inclusion of all essential information, as indicated by the consistent five-star ratings in all evaluated criteria, including "Summary of the Claim" and "Summary of the Defense."

Version V3 provided a comprehensive and well-structured contextualization of facts, offering a cohesive narrative with clear and detailed accounts of procedural occurrences and evidence. While other versions performed well, they lacked the same level of detail and cohesion. Version V3 also excelled in detailing the legal backgrounds in both the initial petition and the contestation, which is crucial for preparing the magistrate's decision. Additionally, V3 followed a clear and precise chronological order, ensuring fluidity and understanding. In summary, V3 was the most complete and well-structured version, meeting the evaluation criteria with the highest clarity, detail, and comprehensiveness. Despite being tested in a specific context of civil courts in Ceará, V3 was recognized for its superior performance. No hallucinations were detected, but human review remains essential for such tasks.

5 Conclusion

The integration of prompt engineering principles, such as the Chain of Density and the CO-STAR framework, along with domain knowledge, has led to the creation of more comprehensible and useful summaries for legal experts. These methods emphasize the importance of essential elements in the process report and the manner, especially the order, in which the report should be generated, effectively meeting the needs of specialists in the legal field.

The combination of generic in-context learning techniques with domain-specific knowledge has proven essential for automatically generating the process report. This application has significant social implications, as it reliably accelerates a crucial stage of the judicial process. Although we focused on a specific type of summary, the generic nature of the process report, which involves various elements, suggests that our approach is broadly applicable to other types of legal document summaries. Further investigation in this area is warranted.

Quantitative metrics for evaluating legal summaries remain limited. Expert-created questionnaires and large-scale validations using GPT models present a promising alternative for the qualitative and quantitative evaluation of different approaches.

Finally, it is worth highlighting the practical viability of SARA, which is already in use by magistrates in a major Brazilian court, significantly contributing to the streamlining of National Justice. The deployment of SARA in a real-world judicial setting underscores its robustness and adaptability to the complex requirements of legal processes. By automating the generation of process reports, SARA not only reduces the time and effort required by legal professionals but also enhances the consistency and accuracy of the reports, ensuring that essential details are systematically captured and presented.

The successful implementation of SARA demonstrates its potential to be scaled and adapted for use in other courts across different jurisdictions, further amplifying its impact on the judicial system. As courts globally face increasing caseloads and the pressure to expedite legal proceedings, tools like SARA offer a viable solution to these challenges by providing a reliable, efficient, and standardized method for handling procedural documentation.

As SARA continues to be refined and expanded, its role in the judicial process could extend beyond report generation to encompass other aspects of legal document management and analysis. The ongoing development and integration of such AI-driven tools represent a transformative step toward a more modern, efficient, and accessible judicial system, ultimately benefiting legal professionals, litigants, and society as a whole.

References

1. OPENAI. Hello GPT-4o. Disponível em: `https://openai.com/index/hello-gpt-4o/`. Acesso em: 30 jun. 2024.
2. ADAMS, Griffin; FABBRI, Alex; LADHAK, Faisal; LEHMAN, Eric; ELHADAD, Noémie. From Sparse to Dense: GPT-4 Summarization with Chain of Density Prompting. In: *Proceedings of the 4th New Frontiers in Summarization Workshop*. Association for Computational Linguistics, 2023. Disponível em: `https://aclanthology.org/2023.propor-1.33`. Acesso em: 1 set. 2024.
3. GUNAWAN, Adrian. How I Won Singapore's GPT-4 Prompt Engineering Competition. Disponível em: `https://towardsdatascience.com/how-i-won-singapores-gpt-4-prompt-engineering-\protect\penalty-\@Mcompetition-34c195a93d41`. Acesso em: 30 jun. 2024.
4. DIAS, P.; MENDES, A.; MARTINS, A. F. T. Improving abstractive summarization with energy-based re-ranking. arXiv, 2022. doi: 10.48550/arXiv.2210.15553.
5. IBM. Engenharia de prompts: a arte de fazer perguntas a IA. IBM, 2023. Disponível em: `https://www.ibm.com/br-pt/topics/prompt-engineering`. Acesso em: 24 jun. 2024.
6. LIU, Y.; ITER, D.; XU, Y.; WANG, S.; XU, R.; ZHU, C. G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment. arXiv, 2023. doi: 10.48550/arXiv.2303.16634.

7. HEILMANN, C. M. Text Summarization of Legal Documents Using Reinforcement Learning: A Study. pp. 403–414, 2021. doi: 10.1007/978-981-19-2894-9_30.
8. KANAPALA, A.; PAL, S.; PAMULA, R. Text summarization from legal documents: a survey. *Artificial Intelligence Review*, 51(3):371–402, 2019. doi: 10.1007/s10462-017-9566-2.
9. ELARABY, M.; LITMAN, D. J. ArgLegalSumm: Improving Abstractive Summarization of Legal Documents with Argument Mining. arXiv, 2022. doi: 10.48550/arXiv.2209.01650.
10. BHATTACHARYA, P.; PODDAR, S.; RUDRA, K.; GHOSH, K.; GHOSH, S. Incorporating Domain Knowledge for Extractive Summarization of Legal Case Documents. arXiv, 2021. Disponível em: `https://arxiv.org/abs/2209.01650`. Acesso em: 30 jun. 2024.
11. KANAPALA, A.; JANNU, S.; PAMULA, R. Summarization of legal judgments using gravitational search algorithm. *Neural Computing and Applications*, Springer Nature, 2019.
12. CONSELHO NACIONAL DE JUSTIÇA (CNJ). Automação processual amplia eficiência de atendimentos em tribunal. Disponível em: `https://www.cnj.jus.br/revista-cnj-automacao-processual-amplia-eficiencia-de-atendimentos/break-em-tribunal/`. Acesso em: 30 jun. 2024.
13. FEIJÓ, Diego de Vargas. Summarizing Legal Rulings. Universidade Federal do Rio Grande do Sul, Instituto de Informática, Programa de Pós-Graduação em Computação. Porto Alegre, 2021.
14. LINS, Alex Aguiar; CARVALHO, Cecília; BOMFIM, Francisco das Chagas Jucá; BENTES, Daniel de Carvalho; PINHEIRO, Vládia. CLSJUR.BR - A Model for Abstractive Summarization of Legal Documents in Portuguese Language based on Contrastive Learning. In: *Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 1*, Santiago de Compostela, Galicia/Espanha, março 2024. Association for Computational Linguistics. Disponível em: `https://aclanthology.org/2024.propor-1.33`.
15. LIU, Y.; LIU, P. SimCLS: A Simple Framework for Contrastive Learning of Abstractive Summarization. arXiv, 2021. Disponível em: `https://arxiv.org/abs/2106.01890`. Acesso em: 1 jul. 2024.
16. VASWANI, A.; SHAZEER, N.; PARMAR, N.; USZKOREIT, J.; JONES, L.; GÓMEZ, A. N.; KAISER, Ł.; POLOSUKHIN, I. Attention is all you need. In: *Advances in Neural Information Processing Systems*, 2017. pp. 5998–6008.
17. BRASIL. Supremo Tribunal Federal. Chamamento Público para Pesquisa sobre IA no STF. Disponível em: `https://www.stf.jus.br/arquivo/cms/noticiaNoticiaStf/anexo/ChamamentoPblicoIASTF.pdf`. Acesso em: 1 jul. 2024.
18. LIU, Y.; LAPATA, M. mBART: Multilingual denoising pre-training for neural machine translation. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020. pp. 7871–7880.
19. SILVEIRA, R.; PONTE, C.; ALMEIDA, V.; PINHEIRO, V.; FURTADO, V. LegalBERT-pt: A Pretrained Language Model for the Brazilian Portuguese Legal Domain. In: *Intelligent Systems*. BRACIS 2023. Lecture Notes in Computer Science.
20. LIN, Chin-Yew. ROUGE: A Package for Automatic Evaluation of Summaries. In: *Proceedings of the Workshop on Text Summarization Branches Out (WAS 2004)*. Barcelona, Espanha: Association for Computational Linguistics, 2004. pp. 25–26. Disponível em: `https://aclanthology.org/W04-1013`. Acesso em: 1 jul. 2024.

21. PAPINENI, Kishore; ROUKOS, Salim; WARD, Todd; ZHU, Wei-Jing. BLEU: A Method for Automatic Evaluation of Machine Translation. In: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*. Philadelphia, PA: Association for Computational Linguistics, 2002. pp. 311–318. Disponível em: `https://aclanthology.org/P02-1040`. Acesso em: 1 jul. 2024.