Information Extraction from Brazilian News Articles on Public Procurement Fraud

Paulo Marcos de Assis[1], **Márcio Castro**[1], **Jônata Tyska Carvalho**[1]

[1] Department of Informatics and Statistics – Graduate Program in Computer Science
Federal University of Santa Catarina (UFSC)

paulo.marcos@grad.ufsc.br, {marcio.castro, jonata.tyska}@ufsc.br

Abstract. *Combating fraud in public procurement is a critical task for oversight agencies. While journalistic coverage can offers a rich source of early warning signals, its unstructured nature hinders automated auditing. To bridge this gap, we propose an Information Extraction (IE) pipeline to extract four key entities required for administrative linkage: municipality, procurement modality, notice number, and contract object. We evaluate three extraction approaches: (1) a Hybrid-Heuristic baseline; (2) a Retrieval-Augmented Generation (RAG) framework; and (3) an End-to-End Structured Large Language Model (LLM). Results demonstrate that the structured LLM significantly outperforms modular baselines, achieving an average accuracy of 91.45% across the four attributes. Furthermore, we introduce a gold-standard dataset of 796 annotated articles. All experiments are reproducible.*

1. Introduction

Journalistic coverage is an essential source of detection in cases of corruption [OECD 2018]. In many countries, the media exposes cases of irregularities that can be the catalyst for investigations by monitoring agencies [UNODC 2014]. There are many examples of journalistic reports that served as the basis for investigations that resulted in the recovery of millions or even billions to the public purse. In 2016, the famous Panama Papers case began after allegations made by a group of 376 journalists from around the world, who exposed a network of financial fraud and political corruption. With the conclusion of the investigations, US$1.2 billion was recovered for the public coffers.

For investigative agencies, access to information that can reveal potential fraud is of great importance [Kariuki and Reddy 2017]. In cases of fraud in public procurement, journalistic coverage serves as an essential data source, providing early indications of irregularities and playing a key role in supporting and initiating official investigations. A notable case in which media reports led to official scrutiny occurred in the Brazilian state of Santa Catarina during the early months of the COVID-19 pandemic. The Court of Auditors of Santa Catarina (TCE/SC) initiated investigations [1] following reports by media outlets such as The Intercept Brasil[2], NSC Total[3], and ND Mais[4]. These reports exposed irregularities in the emergency purchase of 200 pulmonary ventilators for R$ 33 million, procured without a bid process and never delivered to the state.

[1] mpc.sc.gov.br
[2] intercept.com.br
[3] nsctotal.com.br
[4] ndmais.com.br

Despite the relevance of journalistic reports, their use as a source of information for fraud detection presents significant challenges. News articles are produced in large numbers daily and scattered across multiple news portals, making systematic monitoring difficult. In this sense, although early signals of irregularities may be available in the media, investigative agencies still lack automated systems capable of transforming this information into structured data that can support auditing and cross-referencing with official records. As a result, this process remains complex and largely manual.

Recent advances in Machine Learning (ML) and Natural Language Processing (NLP) have enabled the development of automated pipelines to process large volumes of textual data [Schneider dos Santos et al. 2025, Premasiri et al. 2025]. In our previous work, we successfully addressed the problem of identifying relevant news articles, with classification models capable of filtering procurement fraud reports from the daily news stream [Assis et al. 2025]. However, identifying these articles represents only a partial solution. Operating directly on the output of this prior classification stage, our present work addresses the next critical bottleneck: although relevant reports are successfully identified, the information they contain remains unstructured, preventing direct integration with official procurement databases. Transforming unstructured journalistic narratives into structured data is a non-trivial challenge. Unlike official documents, news articles are characterized by high linguistic variability, implicit references, and context-dependent expressions. These characteristics pose significant challenges for traditional Information Extraction (IE) approaches and emerging Large Language Model (LLM)-based methods.

To address this challenge, multiple IE strategies can be considered, ranging from rule-based pipelines and retrieval-augmented methods to fully generative approaches [de Aquino et al. 2024, Hindi et al. 2025, Wei et al. 2024]. However, there is currently limited empirical evidence on how these strategies compare when applied to unstructured news data in the context of public procurement fraud. In this work, we investigate this problem through a comparative evaluation of three distinct IE approaches: (1) a Hybrid-Heuristic pipeline combining Regex and dictionary lookup; (2) a Retrieval-Augmented Generation (RAG) framework with semantic retrieval; and (3) an end-to-end Structured LLM approach. These specific methods were selected to represent a progressive capacity of NLP capabilities: from computationally lightweight syntactic matching (Hybrid-Heuristic), to intermediate chunk-based semantic retrieval (RAG), and finally to full-context generative comprehension (Structured LLM). Our objective is to determine which strategy is most effective for extracting procurement-related attributes, namely *municipality*, *procurement modality*, *notice number*, and *contract object*, from Brazilian news articles.

This work makes three main contributions: (i) a comparative evaluation of IE strategies for structuring data from journalistic reports of procurement fraud; (ii) empirical evidence demonstrating the superiority of structured LLMs over modular pipelines in this domain; and (iii) a gold-standard dataset of 796 annotated news articles, enabling future benchmarking and research. Our results demonstrate that the structured LLM approach significantly outperforms both heuristic and RAG-based pipelines, achieving an overall strict match accuracy of 91.45% on a manually annotated blind test set. The findings indicate that preserving full-document context is crucial for resolving domain-specific ambiguities, particularly in cases involving polysemous terms and implicit references.

The remainder of this paper is organized as follows: Section 2 reviews related work. Section 3 describes the proposed extraction architectures and experimental setup. Section 4 presents the results and discussion. Finally, Section 5 concludes the paper and outlines directions for future work.

2. Related Work

Information Extraction and fraud detection in public procurement have become critical research fronts, particularly given that corruption can inflate contract values by up to 20% [Schneider dos Santos et al. 2025]. Core IE tasks, such as Named Entity Recognition (NER) and Event Detection, are fundamental for transforming unstructured legal text into structured data [Premasiri et al. 2025]. In the Brazilian context, prior work has largely relied on discriminative approaches, including the use of BERTimbau to identify fraud-related "red flags" in procurement documents [Lima et al. 2023] and BERT-based ensembles for entity extraction from the Official Federal Gazette (DOU) [Sirqueira and Vidal 2024].

More recently, research has shifted toward generative paradigms. Retrieval-Augmented Generation has been widely adopted to improve the faithfulness of Large Language Models in high-stakes legal settings [Hindi et al. 2025]. Approaches such as ChatIE [Wei et al. 2024] reformulate IE as a multi-turn question answering task, enabling zero-shot extraction, while systems like INACIA [Pereira et al. 2025] and Lawbot [Solanki et al. 2025] leverage RAG pipelines to support legal decision-making and improve access to legal information. In Brazil, generative models have also demonstrated strong performance in structured information extraction from legal documents [de Aquino et al. 2024] and in simplifying complex legal texts [Silva et al. 2024].

Despite these advances, existing approaches are predominantly evaluated on structured or semi-structured sources, such as official documents and public records. In contrast, journalistic reports of procurement fraud remain underexplored, despite being a valuable source of early signals of irregularities. Unlike standardized administrative texts, news articles exhibit high linguistic variability, implicit references, and domain-specific ambiguity, posing challenges for both traditional IE pipelines and current LLM-based approaches. While efforts such as the CONO crawler [Souza and Dorneles 2025] automate the collection of news related to corruption, they are limited to shallow metadata extraction and keyword-based filtering, lacking mechanisms for domain-specific structured extraction. [Assis et al. 2025] developed a specialized text classification pipeline, demonstrating that context-sensitive models could achieve near-perfect recall in flagging procurement fraud reports in the news. Although [Assis et al. 2025] effectively addresses the classification step by filtering a high volume of daily news, the resulting text data remain unstructured and unsuitable for direct cross-referencing in databases.

The knowledge of how different IE strategies perform when applied to unstructured journalistic narratives in the context of public procurement fraud remains limited. In particular, it remains unclear whether modular approaches, such as heuristic pipelines or RAG-based architectures, are sufficient or whether end-to-end structured LLMs provide a more effective alternative. To address this gap, we conduct a comparative evaluation of three distinct IE strategies for extracting procurement-related attributes from Brazilian news articles: (i) a Hybrid-Heuristic pipeline combining Regex and dictionary lookup;

(ii) a Retrieval-Augmented Generation (RAG) approach with semantic retrieval; and (iii) an end-to-end Structured LLM. By systematically analyzing their performance on a manually annotated dataset, this study provides empirical evidence of the trade-offs among syntactic, retrieval-based, and fully generative approaches to Information Extraction in unstructured journalistic data.

3. Methodology

This section details the methodological approach developed to extract structured attributes from unstructured news texts. We first define the scope of the extraction task and the target attributes required for auditing applications. Subsequently, we present the three distinct strategies investigated in this study, ranging from heuristic baselines to end-to-end generative models. Finally, we outline the experimental setup, including the ground truth creation process, our dataset, the evaluation protocols used to assess performance, and the hardware infrastructure used to validate the proposed pipelines.

3.1. Task Definition and Input Scope

The extraction module operates on the output of a prior classification stage that identifies news articles reporting public procurement fraud. This stage, introduced in [Assis et al. 2025], consists of a supervised text classification pipeline trained on annotated news articles labeled as fraud and non-fraud. The classification model combines contextual embeddings (BERTimbau [Souza et al. 2020]) with a Support Vector Machine (SVM) algorithm to filter relevant cases from the daily news stream with high precision and recall.

After the classification stage, the extraction module operates on the subset of news articles explicitly classified as "public procurement fraud". To enable automated identification of the specific bidding process mentioned in the news, the pipeline targets a quadruple of canonical attributes: the **municipality** (the specific public entity responsible for issuing the tender); the **procurement modality** (the legal category under which the contract was processed); the **notice number** (the alphanumeric registration of the official document); and the **contract object** (the textual description of the contracted goods, works, or services). Although the simultaneous presence of all four attributes within a single news article is empirically rare, the extraction of the maximal available subset is an important objective. Recovering the most complete tuple possible is essential to minimize search ambiguity, thereby maximizing the probability of successfully cross-referencing the reported irregularity against official government databases.

3.2. Extraction Architectures

To determine the optimal strategy for converting unstructured news texts into the target structural format, this study investigates three distinct architectural paradigms, progressing from syntactic matching to semantic understanding.

3.2.1. Hybrid-Heuristic Approach

This first approach decomposes the extraction task based on the linguistic complexity of each attribute. For the *notice number*, strictly defined Regular Expressions (Regex) are

implemented. For *municipality* and *modality* entities, which belong to finite domains, our pipeline utilizes a dictionary-based lookup system indexed via the Apache Solr search engine. Generative extraction via LLM is reserved exclusively for the *contract object* attribute, as this field requires semantic summarization capabilities to handle the high variability of narrative descriptions that heuristics cannot provide.

During the error analysis phase on the development set, this syntactic approach exhibited critical limitations regarding the *modality* attribute. Terms such as "concorrência" (competition) and "concurso" (contest) are polysemous in Portuguese; they frequently appear in news contexts unrelated to procurement law (e.g., *"frustrar a concorrência de mercado"* or *"concurso público de pessoal"*). These linguistic ambiguities resulted in a high rate of false positives, demonstrating that syntactic matching was insufficient for context-dependent entities in our scenario, thereby motivating the shift to context-aware architectures.

3.2.2. Retrieval-Augmented Generation (RAG)

The second approach implements an RAG framework, designed to address the computational cost of processing full-text documents. In this architecture, the input text is segmented into overlapping chunks (small fragments) to maximize the permanence of local context. The Facebook AI Similarity Search (FAISS) is utilized as a vector store to index these segments for semantic retrieval. For each attribute, the system retrieves only the top-k most relevant text chunks (where $k = 2$), which are then fed into the Large Language Model ($gpt - oss : 20b$) for disambiguation.

3.2.3. Structured Large Language Model

The third methodology, which yielded the superior results presented in Section 4, is the end-to-end structured LLM. This architecture leverages the full text of each news article and a prompt engineering strategy using the $gpt - oss : 20b$ model to extract the four attributes as a simultaneous task. The method employs a structured prompt that integrates three core components:

1. **Domain-Specific Constraints:** These rules impose strict constraints on the model, requiring that all attributes be extracted exclusively from the information explicitly present in the text. This includes normalization logic, such as identifying the principal municipality when multiple locations are cited and filtering out generic terms that do not qualify as licitation modalities.
2. **Few-Shot Reasoning:** The prompt embeds examples presented within the context window to resolve ambiguity. These demonstrations illustrate, for instance, how to distinguish "Concorrência" as a modality from market competition contexts, and how to isolate the exact textual excerpt describing the procurement object.
3. **Schema Enforcement:** The output is constrained to a strictly valid JSON object with fixed attribute arrays (*municipality*, *modality*, *notice number*, and *contract object*).

An overview of the pipeline is illustrated in Figure 1. The process starts from a corpus of news articles classified as procurement fraud and compares three extraction

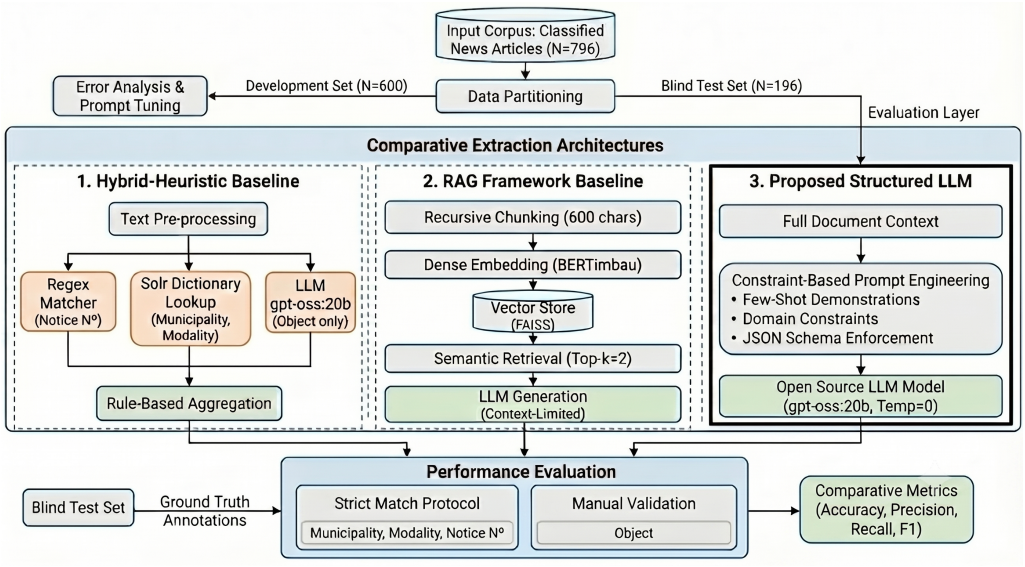

Figure 1. Overview of the comparative Information Extraction architectures

strategies: a hybrid-heuristic baseline, a RAG-based approach, and an end-to-end structured LLM.

3.2.4. Datasets and Partitioning

As one of the contributions of this work, we introduce a gold-standard dataset comprising 796 Portuguese news articles exclusively related to public procurement fraud cases in the state of Santa Catarina, Brazil. To support robust evaluation and ensure generalization capability, the manually annotated ground truth was partitioned into two distinct subsets:

1. **Development Set ($N = 600$):** Representing approximately 80% of the data, this subset was utilized for iterative refinement of the extraction logic and prompt tuning. Specifically, it supported the error analysis phase, which revealed challenges related to polysemous terms in Brazilian Portuguese, such as "Concorrência" and "Concurso", which may assume different meanings depending on the legal and administrative context. This ambiguity motivated the transition to generative models.
2. **Test Set ($N = 196$):** The remaining 20% was preserved as a strictly hold-out set. No prompt engineering or heuristic tuning was performed on these articles. This set was used exclusively for the final performance assessment.

3.3. Experimental Setup

Each news article in the corpus was manually analyzed and annotated to build a structured ground truth. For each of the four target attributes (*municipality*, *modality*, *notice number* and *contract object*), two distinct control columns were annotated:

1. **Content Value:** The exact textual span representing the attribute (e.g., *"'Itajaí"* for *municipality*). If the attribute was absent in the text, this field remained empty.
2. **Presence Indicator:** A binary flag explicitly marking the existence of the attribute in the narrative (1 for present, 0 for absent).

Table 1. Experimental Configuration and Hyperparameters

Component	Configuration
Model	`gpt-oss:20b` (via Ollama)
Temperature	0.0
Context Window	128K tokens
Embedding Model	`neuralmind/bert-base-portuguese-cased`
Chunk Size	600
Chunk Overlap	100
Prompt Strategy	Few-Shot
Retrieval (k)	Top-2 chunks

Table 2. Hardware and Infrastructure Specifications

Parameter	Configuration
GPU	NVIDIA RTX A6000 (48 GB VRAM)
CPU	AMD EPYC 9654 (96-Core)
System RAM	1.5 TB
OS / Drivers	Linux (Kernel 6.8) / CUDA 12.8

This dual-column structure ensures that the evaluation metric distinguishes between a model correctly identifying an absence (true negative) and failing to identify a presence (false negative). The classification of each extraction attempt is determined by comparing the system's output list against the ground truth tuple, according to the following logic:

- **True Positive (TP):** The attribute is present, and the system's extracted list strictly matches the annotated value.
- **True Negative (TN):** The attribute is explicitly absent, and the system correctly returns an empty list.
- **False Positive (FP):** The attribute is absent, but the system hallucinates a value; OR the attribute is present, but the system extracts an incorrect value.
- **False Negative (FN):** The attribute is present, but the system fails to retrieve it, returning an empty list.

This logic was codified into the validation scripts for all three different approaches.

3.3.1. Implementation Details and Reproducibility

All experiments were conducted in a controlled environment using the Ollama framework [5] for model orchestration. The specific hyperparameters and configuration settings are detailed in Table 1, and all hardware infrastructure is detailed in Table 2. The code and datasets used in this work are publicly available[6].

[5] ollama.com

[6] github.com/Paulo-Marcos-Assis/comparative-information-extraction

All experiments were executed on an infrastructure equipped with an AMD EPYC 9654 96-Core Processor and 1.5 TB of RAM. For Large Language Model inference, we used an NVIDIA RTX A6000 GPU with 48 GB of VRAM, running on CUDA 12.8. This specification ensures that the 20-billion-parameter model (`gpt-oss:20b`) fits entirely within GPU memory.

4. Results and Discussions

This section presents the performance evaluation of the proposed IE architectures. The results reported here are exclusively from the test set ($N = 196$), ensuring that the metrics reflect the models generalization capability on unseen data. Performance is summarized using accuracy, precision, recall, and F1-score metrics calculated across the blind test set. To provide a comprehensive comparison, Table 3 presents the detailed performance metrics by attribute for all evaluated extraction approaches. These granular results further highlight the generative approach's superiority in handling contextual filtering and resolving polysemy compared to the modular baselines.

Analyzing the trade-off between precision and recall reveals the fundamental limitation of the Hybrid-Heuristic baseline. Although it achieved the highest recall (94.67%), indicating that the Regex and Solr dictionaries rarely miss a relevant entity, its precision was notably low (56.03%). A granular analysis demonstrates that this performance penalty is primarily driven by the *municipality* attribute, which alone generated 119 false positives within the 196 test records. This high error rate occurs because the Solr dictionary lookup does not consider the surrounding context. News articles on procurement fraud frequently cite multiple municipalities to establish a regional context or describe the scope of investigative operations. However, the objective of this pipeline is to enable future cross-referencing with official databases, which requires precisely isolating the contracting authority in question rather than all geographic mentions. Under the Strict Match evaluation protocol, which requires the extracted set to be identical to the ground truth, the inclusion of these auxiliary municipalities results in the entire prediction being classified as a false positive.

The Structured LLM approach demonstrated substantial superiority, achieving a general accuracy of 91.45%. This represents an improvement of 13.9 percentage points over the Hybrid baseline (77.54%) and 15.2 points over the RAG architecture (76.27%). Due to the descriptive nature of the *contract object* attribute, its extraction was manually validated to avoid unfair penalties from strict-match heuristics. Under this protocol, the RAG architecture achieved an accuracy of 69.90% and an F1-score of 78.85%. While effective, it still lagged behind the Structured LLM, which achieved an F1-score of 96.05%. The performance gap (17% in F1-score) suggests that while RAG successfully retrieves relevant chunks, narrative fragmentation still results in the loss of critical descriptive details that the full-context model can capture. This disparity supports the hypothesis that simultaneous full-context inference via a prompt engineering strategy is more effective in this domain than fragmented retrieval (RAG) or syntactic matching (Hybrid).

The experimental results provide strong evidence that schema-constrained generative models outperform traditional modular pipelines for this specific Information Extraction (IE) task. The performance on *notice number* extraction ($accuracy = 98.47\%$) demonstrates that LLMs can match the precision of strict regular expressions for rigid al-

Table 3. Detailed performance metrics by attribute for all evaluated extraction approaches.

Approach	Attribute	Accuracy	Precision	Recall	F1-Score
1. Hybrid-Heuristic	Municipality	36.22%	28.31%	88.68%	42.92%
	Modality	89.29%	72.58%	91.84%	81.08%
	Notice Number	91.33%	29.17%	100.00%	45.16%
	Object (manual)	93.33%	94.05%	98.14%	96.05%
	Global Average	77.54%	56.03%	**94.67%**	66.30%
2. RAG Baseline	Municipality	58.16%	65.55%	65.55%	65.55%
	Modality	80.61%	51.02%	64.10%	56.82%
	Notice Number	96.43%	75.00%	69.23%	72.00%
	Object (manual)	69.90%	80.87%	76.92%	78.85%
	Global Average	76.27%	68.11%	68.95%	68.30%
3. Structured LLM	Municipality	80.61%	82.89%	91.30%	86.90%
	Modality	93.37%	90.20%	85.19%	87.62%
	Notice Number	98.47%	85.71%	92.31%	88.89%
	Object (manual)	93.33%	94.05%	98.14%	96.05%
	Global Average	**91.45%**	**88.21%**	91.73%	**89.86%**

phanumeric patterns. Simultaneously, the generative approach surpasses heuristic methods in semantically complex fields such as *modality* and *contract object*, where contextual understanding is required to resolve polysemy. Regarding the architectural trade-offs, the reduced performance of the RAG baseline suggests that for news articles of this typical length, maintaining the global narrative structure is more beneficial than optimizing for token retrieval efficiency. The fragmentation inherent to the chunking process appears to sever critical semantic links required to fully characterize the procurement object, a limitation that the full-context Structured LLM effectively overcomes.

However, this superior accuracy comes at a higher computational cost. Running full-document inference with a 20-billion-parameter model requires significant hardware resources (e.g., Table 2), as detailed in our experimental setup. From a practical scalability perspective, processing the entire daily volume of published news with this generative architecture (LLM's) would be cost-prohibitive for most public auditing agencies. For instance, the two largest news portals in the Brazilian state of Santa Catarina (*NSC Total* and *ND Mais*) collectively publish over 500 articles daily. This sheer volume highlights the critical importance of our two-stage pipeline, which uses our previously presented text classification work [Assis et al. 2025] as its first stage. This prior classification acts as an efficient filter, discarding irrelevant news and forwarding only a small subset of high-probability fraud reports to the present extraction module. By restricting the computationally intensive LLM Structured extraction exclusively to this reduced volume, the proposed system ensures that the extraction remains economically viable and scalable for daily real-world monitoring.

Beyond technical performance, extracting the target attributes (*municipality*,

modality, *notice number*, and *contract object*) is a fundamental requirement for automated database linkage. While a complete tuple enables direct, deterministic matching against official procurement records, our analysis confirms that identifiers such as the *notice number* are frequently omitted from journalistic narratives. This data sparsity underscores that the real-world utility of news-based auditing depends on the high-precision extraction of the available subset of attributes. By providing reliable, structured data for the *municipality* and *contract object* attributes, the proposed pipeline identifies the research and implementation of non-deterministic matching strategies as a natural next step for effectively linking media-reported fraud to public databases, even when formal registration numbers are absent from the source text.

Furthermore, regarding the generalizability of these findings beyond the current dataset, the proposed Structured LLM architecture demonstrates potential for broader applicability. While our ground truth is currently constrained to cases in Santa Catarina, public procurement in Brazil is governed by uniform federal legislation, ensuring that the core terminology surrounding modalities and fraud mechanisms remains consistent nationwide. Because the generative approach relies on semantic comprehension and in-context learning rather than region-specific rules, it is inherently equipped to handle different regional journalistic styles. Consequently, we anticipate that this pipeline will maintain robust extraction capabilities when applied to news corpora from other Brazilian states, requiring minimal prompt adaptation.

5. Conclusion and Future Work

This study evaluated three IE architectures to bridge the gap between news classification and structured procurement auditing. Our findings demonstrate that a full-context Structured LLM approach significantly outperforms heuristic and retrieval-augmented baselines, achieving an overall accuracy of 91.45%. By preserving narrative cohesion, this architecture successfully resolves domain-specific ambiguities and extracts complex descriptive attributes that modular pipelines fail to capture.

Despite these advancements, the computational cost of full-document inference remains a challenge. Consequently, future research should pursue four strategic directions: (1) efficiency via optimized retrieval: we aim to refine the RAG-based pipeline by improving chunking and candidate ranking strategies. The goal is to reduce token consumption without compromising the semantic integrity required to characterize procurement objects; (2) advanced database linkage and reconciliation: the ultimate goal is to operationalize the extracted tuples against official repositories. This includes national platforms like the *Portal Nacional de Contratações Públicas* (PNCP), as well as state-level systems such as the Santa Catarina State Court of Accounts' integrated audit system (e-Sfinge) and the Official Gazette of Santa Catarina Municipalities (DOM/SC). Given the identified scarcity of formal notice numbers in journalistic media, this module will explore probabilistic record linkage and semantic matching techniques. This integration will enable public prosecutors to automatically cross-reference media reports with official records, providing a robust tool for data-driven investigative validation; and (3) comprehensive model ablation studies: since our current generative baseline relies on a single model configuration (using $gpt - oss : 20b$), future evaluations can incorporate a diverse set of Large Language Models. This ablation analysis will be useful for assessing the adaptability of our structured prompt architecture across different model sizes and for de-

termining the optimal trade-off between extraction accuracy and inference costs. Finally, the test set size is limited by the inherent scarcity of news related to public procurement fraud. In this sense, future work (4) could involve expanding the scope to other Brazilian states, which would naturally allow for a larger dataset and provide an opportunity to evaluate the approach's generalizability across different regional journalistic styles and administrative contexts.

6. Use of Generative AI tools

Generative AI tools were used to assist with text structuring, linguistic revision, LaTeX formatting, and image generation during the preparation of this paper. The authors assume full responsibility for the authorship, content, and integrity of the work.

References

Assis, P. M. d., Castro, M., and Carvalho, J. T. (2025). Machine learning-based classification of portuguese news articles on public procurement fraud. In *Anais da I Escola Regional de Aprendizado de Máquina e Inteligência Artificial da Região Sul (ERAMIA-RS 2025)*, Porto Alegre, RS. Sociedade Brasileira de Computação. Accepted; awaiting publication.

de Aquino, I. V., dos Santos, M. M., Dorneles, C. F., and Carvalho, J. T. (2024). Extracting information from brazilian legal documents with retrieval augmented generation. In *Anais do IX Workshop on Data Science against Corruption in the Public Sector (DSCOPS 2024)*, pages 280–287, Florianópolis, SC, Brasil. SBC.

Hindi, M., Mohammed, L., Maaz, O., and Alwarafy, A. (2025). Enhancing the precision and interpretability of retrieval-augmented generation (rag) in legal technology: A survey. *IEEE Access*, 13:46171–46189.

Kariuki, P. and Reddy, P. (2017). Operationalising an effective monitoring and evaluation system for local government: Considerations for best practice. *African Evaluation Journal*, 5(2).

Lima, W., Lira, R., Paiva, A., Silva, J., and Silva, V. (2023). Methodology for automatic extraction of red flags in public procurement. In *2023 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7, Gold Coast, Australia.

OECD (2018). *The Role of the Media and Investigative Journalism in Combating Corruption*. OECD Publishing, Paris.

Pereira, J., Assumpcao, A., Trecenti, J., Airosa, L., Lente, C., Cléto, J., Dobins, G., Nogueira, R., Mitchell, L., and Lotufo, R. (2025). Inacia: Integrating large language models in brazilian audit courts: Opportunities and challenges. *Digital Government: Research and Practice*, 6(1).

Premasiri, D., Ranasinghe, T., Mitkov, R., et al. (2025). Survey on legal information extraction: current status and open challenges. *Knowledge and Information Systems*, 67:1128–1135.

Schneider dos Santos, E., Machado dos Santos, M., Castro, M., et al. (2025). Detection of fraud in public procurement using data-driven methods: A systematic mapping study. *EPJ Data Science*, 14(52). Published 22 July 2025.

Silva, M., Santos, E., Alves, K., Silva, H., Pedrosa, F., Valencca, G., and Brito, K. (2024). Using generative ai for simplifying official documents in the public accounts domain. In *Anais do XII Latin American Symposium on Digital Government (LAS-DIGOV 2024)*, pages 246–253, Brasília, DF, Brasil. SBC.

Sirqueira, E. J. V. and Vidal, F. d. B. (2024). Evaluation of named entity recognition using ensemble in transformers models for brazilian public texts. In *Anais do XXI Encontro Nacional de Inteligência Artificial e Computacional (ENIAC 2024)*, pages 966–977, Belém, PA, Brasil. SBC.

Solanki, A., Main, H., Mehta, D., Kulkarni, D., and Dalvi, H. (2025). Lawbot: An enhanced legal information retrieval system using rag. In *Anais da 2025 IEEE Silchar Subsection Conference (SILCON)*, pages 1–6, Dimapur, India. IEEE.

Souza, A. C. S. d. and Dorneles, C. F. (2025). Cono: Um coletor automatizado de notícias sobre corrupção em santa catarina. In *Anais da XX Escola Regional de Banco de Dados (ERBD)*, pages 129–132, Florianópolis, SC. Sociedade Brasileira de Computação.

Souza, F., Nogueira, R., and Lotufo, R. (2020). Bertimbau: Pretrained bert models for brazilian portuguese. *Intelligent Systems*, pages 403–417.

UNODC (2014). *Reporting on Corruption: A Resource Tool for Governments and Journalists*. United Nations Office on Drugs and Crime, Vienna.

Wei, X., Cui, X., Cheng, N., Wang, X., Zhang, X., Huang, S., Xie, P., Xu, J., Chen, Y., Zhang, M., Jiang, Y., and Han, W. (2024). Chatie: Zero-shot information extraction via chatting with chatgpt.