A Hybrid RAG Pipeline for Linking Fraud News to Public Procurement Records*

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. *Public procurement is a major target for fraud and mismanagement, and linking journalistic investigations to official procurement records can support automated auditing systems. However, news reports rarely mention identifiers such as tender numbers, making the matching task difficult in large databases with many near-identical descriptions. We present a hybrid Retrieval-Augmented Generation (RAG) pipeline combining structured attribute extraction, SQL pre-filtering, TF-IDF retrieval and Large Language Model (LLM) reranking to link fraud-related news to official procurement processes. Experiments were conducted on 315,052 real procurement records from Santa Catarina (Brazil), using 500 synthetic news-process pairs for development and 214 held-out pairs for final evaluation. Among thirteen retrieval methods, classical TF-IDF outperformed Portuguese neural encoders by more than 57 percentage points in top-50 retrieval coverage. Neural Portuguese encoders (BERTimbau, HeIBERT) were evaluated as zero-shot baselines; the 57-point gap versus TF-IDF reflects untrained embeddings and does not represent their fully optimized potential. The best configuration, TF-IDF top-10 retrieval followed by a single* `gpt-oss:20b` *reranking call, achieved 76.8% development accuracy and 71.5% held-out test accuracy.*

1. Introduction

Public procurement is one of the largest components of government spending and, consequently, a recurrent target for fraud, cartel activity and mismanagement [Lyra et al. 2022]. In this context, investigative journalism acts as a critical mechanism for exposing corruption [OECD 2018]. Media reports frequently reveal irregularities that become the starting point for official investigations and the subsequent recovery of public funds [UNODC 2014]. In the Brazilian state of Santa Catarina, for example, the recent "Operação Mensageiro", one of the largest anti-corruption operations in the state's history, exposed a massive cartel operating in solid waste management and public lighting. Investigations that led to the arrest of over a dozen mayors were heavily supported by continuous coverage from local media outlets [1], which brought to light widespread bid-rigging and systemic corruption. Despite the relevance of journalistic reports, their effective use for fraud detection is hindered by the immense volume and fragmentation of daily news. To illustrate, Santa Catarina's two primary news portals, NSC Total [2] and ND

*Code and reproducibility materials: `https://github.com/Paulo-Marcos-Assis/rag-linkage-paper`

[1]ND Mais coverage and Abrema news.

[2]nsctotal.com.br

Mais[3], collectively published over 11,600 articles in May 2026 alone. This scale makes systematic, manual monitoring unfeasible.

Recent literature has addressed this automated auditing challenge through progressive computational stages. Previous studies have successfully deployed Machine Learning (ML) classifiers to filter the massive daily news stream and isolate articles reporting on public procurement fraud [Assis et al. 2025]. Moreover, because journalistic narratives are inherently unstructured and disconnected from administrative registries, subsequent works [Assis et al. 2026] have introduced dedicated Information Extraction (IE) pipelines. These systems map the unstructured text into standardized tuples, extracting essential attributes such as *municipality*, *procurement modality*, *notice number*, and *contract object*. However, we did not identify any studies in the literature that automatically link news related to possible irregularities in public procurement to official databases. Connecting extracted attributes to structured official repositories, such as the *Portal Nacional de Contratações Públicas* (PNCP) and the state's integrated audit system (e-Sfinge), remains a significant challenge. Because formal bidding notice numbers are frequently absent in media discourse, direct matching is rarely viable. To bridge this gap, we propose a linkage methodology relying exclusively on three available attributes: municipality, procurement modality, and the contract object description. At its core is a Retrieval-Augmented Generation (RAG) framework, utilizing the contract object extracted during the IE phase as the primary semantic query to retrieve and rank matching administrative processes.

This task is a complex instance of record linkage. Procurement descriptions follow rigid legal templates, resulting in distinct processes sharing almost identical wording. For context, the database we target contains 315,052 valid processes from Santa Catarina, of which more than one-fifth (65,033) share a byte-for-byte identical object description. News articles, in turn, are summaries written in a journalistic style and do not reproduce this formal phrasing. A successful linking system must therefore bridge a severe vocabulary gap while remaining precise enough to separate nearly identical processes. To evaluate this system rigorously in the absence of large-scale annotated real-world corpora, we construct a controlled synthetic gold-standard benchmark. We utilized a Large Language Model (LLM), `gpt-oss:20b`, to generate 714 fictional news articles in journalistic style based on real procurement records, splitting them into development and held-out test sets.

This work makes four main contributions: (i) an hybrid pipeline designed specifically for Brazilian procurement data, combining structured extraction, SQL prefiltering, classical textual retrieval, and LLM reranking; (ii) an experimental evaluation validating the pipeline through systematic comparison of thirteen retrieval methods and three reranking strategies; (iii) the identification of equivalence-aware evaluation (duplicate/normalized-object grouping) as a strict requirement for sound measurement in this domain; and (iv) a reproducible synthetic gold-standard benchmark of 714 LLM-generated news-to-process pairs. Our results demonstrate that classical TF-IDF significantly outperforms modern neural embeddings in this context, surpassing them by more than 57 percentage points in candidate retrieval coverage. Furthermore, a single-call LLM reranker successfully bridges the vocabulary gap between journalistic and legal texts, rais-

[3] ndmais.com.br

ing development-set end-to-end linkage accuracy from 69.4% to 76.8% and achieving 71.5% on the held-out test set.

The remainder of this paper is organized as follows: Section 2 reviews related work. Section 3 presents our task definition and our Data. Section 4 describes the proposed linkage architecture and the experimental setup. Section 5 presents the results and discussion. Finally, Section 6 concludes the paper.Recent breakthroughs

2. Related Work

Recent breakthroughs in Natural Language Processing (NLP) have enabled the automated processing of massive textual datasets for governmental oversight [Schneider dos Santos et al. 2025, Premasiri et al. 2025]. For example, automated structures are increasingly used to classify news and automate the extraction of data from texts related to public tenders. [Oliveira et al. 2026]. Furthermore, the highly specialized and complex terminology of public procurement documentation poses a significant barrier to general-purpose NLP models. To address this, recent initiatives have developed domain-specific language models pretrained directly on vast corpora of Brazilian tender notices and contracts, capturing the unique linguistic nuances of this domain [Lima et al. 2026].

The candidate-generation stage of our pipeline draws on classical Information Retrieval (IR). TF-IDF weighting and the probabilistic BM25 framework [Robertson and Zaragoza 2009] remain strong baselines, while dense passage retrieval with BERT-style encoders [Karpukhin et al. 2020] has become the standard neural alternative. Comparative studies such as the BEIR benchmark [Thakur et al. 2021] have shown that neural retrievers do not uniformly dominate lexical methods, particularly out of domain. Our results reinforce this observation in a strongly specialized setting.

A critical obstacle to developing and validating linkage systems in the auditing domain is the scarcity of labeled real-world data, as authentic news rarely includes explicit references to official process identifiers. To overcome this type of scenario, the use of synthetic benchmarks generated by language models has become a well-established practice in NLP and IR evaluation [Wang et al. 2021, Bonifacio et al. 2022]. The scarcity of domain-specific training data in the public procurement and contracting sector has driven the development of specialized pipelines, such as struct2unstruct [Abbas et al. 2026]. In parallel, LLMs have been employed to generate synthetic news describing political events and community violence, allowing researchers to train traditional supervised classifiers with robust performance, eliminating the high costs of manual labeling [Halterman 2025].

Recent literature demonstrates the efficacy of generating synthetic queries and documents to simulate realistic search scenarios, ensuring an unambiguous ground truth while preserving the semantic challenges of the domain [Rahmani et al. 2024]. Furthermore, in contexts involving sensitive data, such as public procurement, privacy-preserving data synthesis techniques enable robust and secure entity resolution [Qin et al. 2022]. Ultimately, the primary goal of constructing this synthetic benchmark is to rigorously evaluate the proposed linkage methodology under controlled and auditable conditions, aligning with modern AI-assisted record linkage frameworks [Ather 2026, Karpov et al. 2026].

The use of LLMs as rerankers has attracted considerable attention. Sequence-to-sequence rankers [Nogueira et al. 2020] and ranking-specific fine-tuning [Zhuang et al. 2023] preceded prompt-based listwise rerankers such as RankGPT

[Sun et al. 2023], which demonstrate that general-purpose LLMs can reorder candidate lists competitively. Our work applies this idea to record linkage and contrasts a listwise single-call reranker with a tournament scheme. The overall design follows the Retrieval-Augmented Generation paradigm [Lewis et al. 2020] and the two-stage retrieve-then-reason pipelines of [Izacard and Grave 2021]. For the Portuguese language, the encoders we evaluate include BERTimbau [Souza et al. 2020] and [Lima et al. 2026], a domain-adapted BERT variant.

3. Task definition, Data and Problems related to domain data

Our core task is to link a news article to its corresponding official procurement record. We start with a database where each administrative process is defined by a unique identifier, a municipality, a procurement modality, and a free-text description of the contract object. From a given news item, our extraction pipeline identifies these same three attributes: the municipality, the modality, and the object description. In our synthetic benchmark, the *gold process* is the real procurement record originally used as the source for generating that news article. Its identifier is therefore the ground-truth answer for evaluation. The system's goal is to use the extracted clues to search the database and identify this official process. We evaluate both the candidate-retrieval stage and the final end-to-end linkage pipeline using complementary metrics and evaluation protocols introduced later in Section 3.1. Retrieval quality is measured through Recall at Rank k (R@k), while the final linkage stage is evaluated through matching accuracy.

The knowledge base consists of the complete set of procurement processes from Santa Catarina, originally comprising 325,781 records. After removing entries whose object description was null or empty (10,729 records), 315,052 valid processes remained. Each record carries, besides the textual object description, the municipality, the procurement modality, the managing unit and the edital number. Due to the scarcity of authentic news related to public bidding processes that would allow for the identification of corresponding processes, building large-scale annotated datasets through manual labeling is extremely difficult and time-consuming. We therefore adopted a controlled approach to generating synthetic data: starting from a dataset with 714 real news items related to bid rigging, 714 real records of public purchases extracted from the database of 315,052 bidding processes, and an LLM (`gpt-oss:20b`, temperature 0.3), we constructed a set of fictitious news items. The model is instructed to insert the process attributes (specifically municipality, procurement modality, and a paraphrased version of the object description) without copying the legal wording of the original bidding document. Each generated article is individually paired with the process from which it originated, providing a fundamental truth for evaluating our proposed approach. Since these fictional articles attribute fraudulent conduct to real procurement records from real municipalities, they are not publicly released; the generation prompts, model configuration and sampling criteria documented here allow the benchmark to be regenerated from the public database.

We split the synthetic data into two disjoint sets that follow exactly the same generation procedure and equivalence rules. A *development set* containing 500 news-process pairs (70% of the dataset) was used for method selection and tuning, including the choice of retriever and reranking strategy. A separate *test set* containing 214 pairs (30%) was held out and used only once for final evaluation, simulating performance on unseen data after all design decisions had been finalized. Together, the 714 pairs cover 127 municipalities

and eight modalities, providing diversity across the domain.

Finally, a problematic characteristic of the domain data is textual redundancy. Among the 315,052 processes, 65,033 (20.6%) have an object description identical to at least one other process, forming 25,248 duplicate groups (Table 1).

Table 1. Distribution of byte-identical duplicate object descriptions among the 315,052 procurement processes

Occurrences per description	Groups	Process IDs
Exactly 2 times	20,462	40,924
Exactly 3 times	2,994	8,982
4 or more times	1,792	15,127
Total	**25,248**	**65,033**

3.1. Evaluation Protocols

The duplication patterns described above introduce an important distinction between two evaluation settings used throughout this work. **Strict evaluation** requires the predicted process identifier to exactly match the gold identifier associated with the news item. This protocol is used for candidate-retrieval benchmarking because it enables direct and reproducible comparison across retrieval methods. All retrieval metrics reported as Recall at Rank k (R@k) and mean rank therefore use strict identifier matching unless explicitly stated otherwise. **Equivalence-aware evaluation**, by contrast, considers a prediction correct not only when the predicted identifier exactly matches the gold process, but also when both belong to the same duplicate or identical-object equivalence group.

This distinction materially affects top-1 style metrics. For example, a retrieval result may fail under strict identifier matching while still selecting a semantically equivalent procurement process. As later shown in Section 5, the TF-IDF top-1 result increases from 60.2% under strict evaluation to 69.4% under equivalence-aware evaluation after duplicate and identical-object normalization are taken into account. Throughout the research, the initial recovery experiments use rigorous evaluation (exact match), while the reclassification and final pipeline evaluation use equivalence-aware evaluation, meaning that predictions are also considered correct when the returned process belongs to the same equivalence group of duplicate or identical objects as the reference process.

To model this domain-specific redundancy, we define equivalence groups over procurement processes. Two processes are considered equivalent when they have a byte-identical object description (a *duplicate* group) or when their descriptions become identical after normalization (an *identical-object* group). Normalization lowercases the text, strips accents, removes trailing punctuation and collapses whitespace. As Table 2 shows, this normalization grows the set of identical-object groups from 1,719 to 2,071 and the processes they cover from 4,712 to 5,627. Under the equivalence-aware evaluation protocol, a prediction is counted as correct if the chosen identifier either exactly matches the gold process or belongs to the same duplicate or identical-object group.

4. Methodology of the Hybrid RAG Linkage Pipeline

An overview of the proposed record-linkage architecture is illustrated in Figure 1. The pipeline processes an unstructured news article through four distinct stages. It begins with

Table 2. Effect of progressive normalization on identical-object detection and its contribution to the pipeline accuracy.

Normalization level	Groups	Processes	Accuracy gain
None	1,719	4,712	—
Trailing punctuation	1,941	5,311	$\approx +3$ pp
Accents + case (final)	**2,071**	**5,627**	$+9.8$ **pp**

an output that comes from the IE stage (Stage 1), where a structured LLM extracts key metadata attributes from the text. These attributes are passed to a metadata-driven SQL pre-filter (Stage 2) that aggressively prunes the database search space. Next, the pipeline performs semantic retrieval (Stage 3) via a vector-based mathematical comparison between the numerical representations of the query and the filtered candidates, evaluating both classical TF-IDF and neural embedding paths. Finally, the top-K ranked candidates are passed to an LLM reranking stage (Stage 4), comparing a multi-call tournament scheme against a single-call prompt configuration to determine the final matching procurement process identifier.

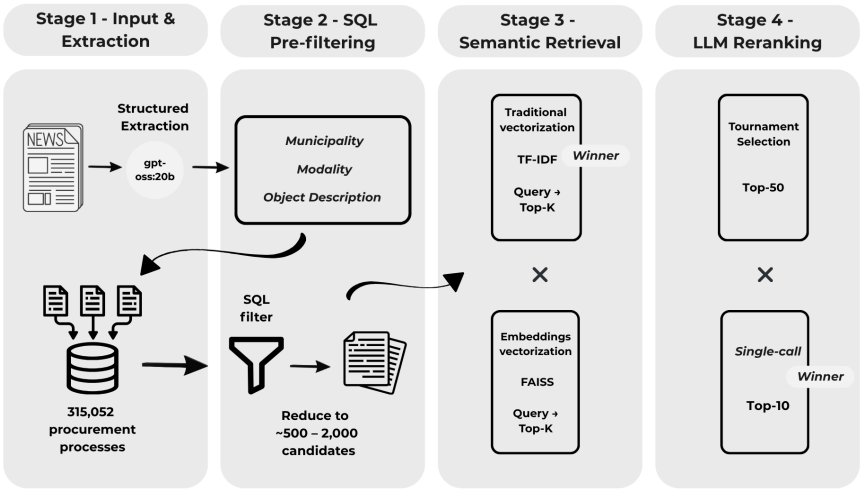


Figure 1. Overview of the proposed hybrid linkage pipeline.

Attribute Extraction and SQL Filtering Stage 1 uses a structured LLM prompt (`gpt-oss:20b`, temperature 0) to extract the municipality, the modality and the object description from the news text. The attributes municipality and modality are then used as exact filters over the database metadata, reducing the search space from 315,052 processes to a few hundred to a few thousand candidates that share both attributes (Stage 2).

Semantic Retrieval of Candidates The filtered candidates are ranked by similarity between the news object and each candidate object. We call this step semantic retrieval (Stage 3), where we look for the top-k best candidates to advance to the next phase of the pipeline. To find the best approach, we evaluated thirteen methods grouped into four families: distance measures (Levenshtein and Jaro-Winkler); set-based token similarities (Jaccard, Dice, and the overlap coefficient); character n-grams with $n \in \{2, 3, 4, 5\}$; TF-IDF cosine similarity with inverse frequency weights of documents estimated across the corpus; and

neural embeddings of three BERT [Devlin et al. 2019] encoders in Portuguese, `bert-base-portuguese-cased`, `bert-large-portuguese-cased` (BERTimbau base and Large [Souza et al. 2020]) and a model pre-trained on bidding documents (HeIBERT [Lima et al. 2026]), indexed with FAISS. All neural encoders were evaluated in a zero-shot setting; that is, no task-specific fine-tuning, contrastive learning, or domain adaptation was applied beyond the pre-trained weights. Therefore, the reported neural results do not represent the maximum performance potential achievable with dense embeddings fine-tuned for the specific task. The role of the retriever is coverage: it must be able to return the gold process within the top-k list that the reclassifier will examine.

LLM Reranking Stage 4 selects one process from the retrieved candidates using an LLM that reasons over the object descriptions. We compare three strategies. The *TF-IDF baseline* performs no reranking and simply returns the top-ranked candidate. The *tournament* strategy splits the top-50 candidates into ten groups of five, asks the model to pick the best in each group and then runs a final round over the ten winners, for about eleven LLM calls per news item; it was run with the smaller `qwen2.5:7b` model. The *single-call* strategy presents the top-10 candidates in one prompt and asks the model to return the index of the matching process, or zero if none matches, in a single call with the larger `gpt-oss:20b` model. All reranking runs use temperature 0, a 300-second timeout and a single attempt per call.

4.1. Experimental Setup

Method-selection experiments were conducted on the 500-item development set against the 315,052-process database, retrieving up to 50 candidates post-SQL filtering. To prevent data leakage, TF-IDF weights were pre-computed exclusively on the official process corpus. LLM inference was served via Ollama at temperature 0 for determinism, employing `qwen2.5:7b` for tournament reranking and `gpt-oss:20b` for the single-call configuration. After optimization, the final pipeline, TF-IDF top-10 retrieval followed by single-call `gpt-oss:20b` reranking, was evaluated once on the strictly disjoint 214-item held-out test set. All experiments ran on an AMD EPYC 9654 CPU (1.5 TB RAM) and an NVIDIA RTX A6000 GPU (48 GB VRAM, CUDA 12.8), allowing the 20B model to fit entirely in VRAM.

5. Results and Discussion

Table 3 reports the retrieval results for the thirteen methods. The most striking finding is the gap between classical textual methods and neural embeddings. TF-IDF with cosine similarity achieves the best overall behaviour, placing the gold process first in 60.2% of the cases (exactly identifier match), within the top five in 87.4% and within the top fifty in 96.0%, with an average rank of 2.9. Character n-grams follow closely; $n = 4$ attains the single highest coverage (96.2% at top-50). In significant contrast, all three neural encoders fail in our task: their average coverage of the 50 main terms is only 37.7%, more than 57 percentage points below TF-IDF, and their R@1 hovers around 20%. It is important to note that this comparison pits an optimized TF-IDF baseline against untrained dense embeddings. The observed 57-point gap reflects the zero-shot performance of the neural encoders; with task-specific fine-tuning or contrastive learning, dense retrievers might narrow or close this gap, but such optimization is outside the scope of this

study. The HeIBERT adapted for tenders performed worst in our experiment (36.6% coverage), indicating that specialization in a tender corpus does not necessarily translate to a test where excerpts from news articles, with journalistic adaptation, are to be compared with descriptions of public tender objects. TF-IDF, in turn, preserved, through high IDF weights, the terms that have great relevance in the comparison between excerpts from news articles and the actual descriptions of the tenders.

This behavior is explained by the nature of the domain, which lies in rare lexical items, street names, product categories, proper nouns and administrative codes, to which TF-IDF assigns high IDF weights precisely because they appear in few of the 315,052 documents. For instance, for the query *"aquisição de materiais para execução de drenagem nas ruas Angelina Motter e Oriente"*, the prepositions *para* and *de* carry IDF weights near 1.2 and contribute little to the final score, whereas *Angelina* (IDF = 8.9), *Motter* (IDF = 11.7) and *Oriente* (IDF = 2.3) dominate the similarity computation. As a result, the gold process reaches a cosine similarity of 0.996 and is returned at rank one, while the second-ranked candidate scores only 0.120.

On the other hand, the embedding-based models often assigned low similarity to pairs that were correct but different in the sense of the journalistic description of an object, compared to the actual description in the document. A representative example is the query *"contratação de empresa para prestação de serviço de seguro total dos equipamentos da usina de asfalto municipal"* (111 characters), the matching process spans 670 characters listing equipment models, voltages and serial numbers (*"Usina de Asfalto Móvel 20-40 Ton/h... Vibroacabadora CIBER AF 4000... Rolo Hamm HD90K..."*). All three coders left the gold candidate out of the top 50, whereas TF-IDF keeps it at rank four. A similar problem also appeared in shorter descriptions. For the query *"contratação de um cesto aéreo isolado de 46 quilovolts, classificado como Classe C"*, the gold object was the compact administrative description *"CESTO AÉREO ISOLADO 46kV - Classe C"*. Although the two texts clearly refer to the same item, all three embedding models again failed to place the gold process in the top-50, whereas TF-IDF returned it in first position. Overall, in 286 of the 500 items (57.2%), all three embedding models left the gold process outside the top-50. Among these same cases, TF-IDF recovered 266 (93.0%) within the top-50 and ranked 164 of them in first position.

The best-performing retrieval approach (TF-IDF) was also run on the test set. Table 4 shows consistent behavior: TF-IDF places the gold process first in 65.9% of cases, within the top ten in 94.4% and within the top fifty in 97.2%, with an average rank of 2.13 and only six gold processes (2.8%) outside the top 50. The agreement between the development and held-out sets confirms that TF-IDF generalizes well as the retrieval component and that a top-10 ranking remains appropriate for the next reranking phase, in an attempt to find the gold candidates.

5.1. LLM Reranking with Single Call Wins

Table 5 compares four reranking configurations. The TF-IDF top-1 baseline, which simply returns the highest-ranked retrieved candidate without reranking, reaches 69.4% accuracy (347/500) - This is distinct from the strict R@1 value in Table 3, which scores exact gold-identifier matches for comparability across retrieval methods. The tournament with the 7B model `qwen2.5:7b` achieves 70.2%. The single-call strategy with the same 7B model, however, drops to 62.4% because the model abstains on 13.8% of items, judging

Table 3. Strict candidate retrieval evaluation on 500 news items against 315,052 processes (top-50), using exact gold-identifier matching.

Method	R@1	R@5	R@10	R@50	Mean rank	Family
TF-IDF cosine	**60.2**	**87.4**	**91.0**	96.0	**2.9**	Textual
N-grams (n=4)	57.2	84.8	89.4	**96.2**	3.5	Textual
N-grams (n=5)	57.4	84.4	89.6	95.8	3.3	Textual
N-grams (n=3)	57.6	83.6	88.2	95.0	3.4	Textual
Overlap coef.	55.2	84.0	87.2	94.4	3.4	Textual
Jaccard	53.8	79.0	84.0	92.8	4.0	Textual
Dice	53.8	79.0	84.0	92.8	4.0	Textual
N-grams (n=2)	51.2	76.2	82.2	90.8	3.9	Textual
Levenshtein	21.0	30.8	35.2	53.6	10.8	Textual
Jaro–Winkler	14.8	25.4	31.4	51.2	12.4	Textual
BERT base	20.6	31.2	34.2	38.0	3.9	Neural
BERTimbau Large	20.4	30.4	34.0	38.6	4.7	Neural
HeIBERT	18.0	28.0	31.8	36.6	5.1	Neural

Table 4. TF-IDF candidate retrieval on the development set (500) versus the held-out test set (214), against 315,052 processes (top-50). R@k is recall at rank k; lower mean rank is better.

Set	R@1	R@5	R@10	R@50	Mean rank
Development (500)	60.2	87.4	91.0	96.0	2.90
Test (214)	**65.9**	**89.7**	**94.4**	**97.2**	**2.13**

that none of the candidates in the round are the gold standard. The single-call strategy with the 20B model `gpt-oss:20b` is the clear winner at 76.8% accuracy, a gain of 7.4 points over the TF-IDF baseline, using one call per item and failing only once in 500 (0.2%). However, this gain comes at a computational cost: while TF-IDF returns results in less than one second per item, the single-call reclassifier 20B requires an average of 20.8 seconds per item.

Table 5. Reranking strategies on the 500-item gold standard. Accuracy is equivalence-aware; latency is the average per news item.

Metric	TF-IDF baseline	Tournament 10×5 (7B)	Single-call top-10 (7B)	Single-call top-10 (20B)
Accuracy	69.4%	70.2%	62.4%	**76.8%**
LLM calls / item	0	11	**1**	**1**
Total LLM calls	0	5,500	500	500
Abstention rate	—	—	13.8%	3.4%
Failure rate	—	3.2%	**0.0%**	0.2%
Latency / item	$<$ 1s	4.2s	53.3s	20.8s

5.2. Final Evaluation on the Held-out Test Set

As a final evaluation, we apply the best configuration identified during the development experiments — retrieval of the top-10 candidates by TF-IDF followed by single-call reranking with `gpt-oss:20b` — to the held-out test set, which did not influence any design decision. This measures how the pipeline generalises to unseen data and pro-

tects against potential biased metrics on the development set. Table 6 reports the results. TF-IDF retrieves an acceptable candidate within the top 10 for 202 of the 214 test items (94.4%), slightly above the corresponding development value of 91.4%. The final pipeline correctly links 153 items, yielding 71.5% accuracy. This number should be interpreted as the accuracy of the complete retrieve and rerank pipeline, not as the isolated accuracy of the LLM. Furthermore, in the 12 cases where the gold process was absent from the top-10 candidate list, no reranker could recover the correct answer.

Table 6. Comparison of results between the final evaluation on the reserved test set and the development set.

Metric	Development (500)	Test (214)
TF-IDF top-10 coverage	91.4% (457/500)	94.4% (202/214)
Single-call accuracy (20B)	76.8% (384/500)	71.5% (153/214)
Abstentions	3.4% (17/500)	2.3% (5/214)
Technical failures	0.2% (1/500)	0.5% (1/214)

The completed 214-item test run produces 61 errors, twelve are irrecoverable by the reranker because the gold process is outside the top-10 candidate list. Among the 202 reachable cases, the model correctly selects 153, for a reachable-case accuracy of 75.7%. The remaining 49 reachable errors therefore reflect reranking or reliability limitations: 48 are incorrect selections among highly similar candidates and one is a technical failure. The five test-set abstentions all occur in cases where the gold process is absent from the top-10 list.

6. Conclusion and Future Works

We presented a hybrid RAG pipeline for linking procurement-fraud news to official bidding records and evaluated it over 315,052 real processes using a 500-item development benchmark plus a 214-item held-out test set. A systematic comparison of thirteen retrieval methods showed that classical TF-IDF surpasses neural Portuguese encoders by more than 57 percentage points of coverage in this domain, and a comparison of three reranking strategies showed that a single call to a 20B model raises development-set accuracy to 76.8%, 7.4 points above the equivalence-aware TF-IDF@1 baseline (69.4%) and 6.6 above a multi-call tournament with a smaller model, while being far more reliable. When this final configuration is evaluated once on the held-out set, it reaches 71.5% accuracy, with the gold process present in the top-10 TF-IDF candidates for 94.4% of test items. We also demonstrated that handling duplicate and normalised-identical descriptions, a structural feature of public procurement, is essential for sound evaluation and accounts for nearly ten points of the development-set accuracy. Future work includes testing the database linking methodology with a set of real news items and studying cases where the retrieval methodology failed to find the correct candidates.

7. 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. No scientific or technical content was generated by AI tools. All the generated content was carefully revised by the authors. The authors assume full responsibility for the authorship, as well as for the content and its integrity.

References

Abbas, A., Mark, L., Shanavas, N., et al. (2026). Struct2unstruct: Creating tender ner datasets from structured procurement records using large language models. *Preprint*.

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.

Assis, P. M. d., Castro, M., and Carvalho, J. T. (2026). Information extraction from brazilian news articles on public procurement fraud. In *Anais do 41º Simpósio Brasileiro de Bancos de Dados (SBBD)*. Accepted for publication; awaiting publication.

Ather, H. (2026). Llm-assisted record linkage: A framework for official statistics. *Statistical Journal of the IAOS*.

Bonifacio, L. H., Abonizio, H., Fadaee, M., and Nogueira, R. (2022). Inpars: Data augmentation for information retrieval using large language models. *ArXiv*, abs/2202.05144.

Devlin, J., Chang, M.-W., Lee, K., and 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 (NAACL-HLT)*.

Halterman, A. (2025). Synthetically generated text for supervised text analysis. *Political Analysis*, 33(3):181–194.

Izacard, G. and Grave, E. (2021). Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 874–880.

Karpov, I., Kirillovich, A., et al. (2026). Synel: A synthetic benchmark for entity linking. *Expert Systems with Applications*.

Karpukhin, V., Oğuz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., and Yih, W.-t. (2020). Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., Riedel, S., and Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 9459–9474.

Lima, W. E. M., da Silva, V. R., da Silva, J. C., Rabêlo, R. d. A. L., and de Paiva, A. C. (2026). Helbert: A bert-based pretraining model for public procurement tasks in portuguese. *Journal of the Brazilian Computer Society*.

Lyra, M. S., Damásio, B., Pinheiro, F. L., et al. (2022). Fraud, corruption, and collusion in public procurement activities, a systematic literature review on data-driven methods. *Applied Network Science*, 7:83.

Nogueira, R., Jiang, Z., Pradeep, R., and Lin, J. (2020). Document ranking with a pre-trained sequence-to-sequence model. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 708–718.

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

Oliveira, E. T. L., Alves, P. M. H., Mendes, F. I. d. L., da Silva Júnior, E. F. P., Alves, A. L. F., and Baptista, C. d. S. (2026). Improving audits in public procurement: An approach to recognition of named entities for brazilian official gazettes. *Procedia Computer Science*, pages 141–148.

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.

Qin, X., Chai, C., Tang, N., et al. (2022). Synthesizing privacy preserving entity resolution datasets. In *2022 IEEE 38th International Conference on Data Engineering*.

Rahmani, H. A., Craswell, N., Yilmaz, E., et al. (2024). Synthetic test collections for retrieval evaluation. *arXiv preprint arXiv:2405.07767*.

Robertson, S. and Zaragoza, H. (2009). The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389.

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.

Souza, F., Nogueira, R., and Lotufo, R. (2020). Bertimbau: Pretrained bert models for brazilian portuguese. In *Proceedings of the 9th Brazilian Conference on Intelligent Systems (BRACIS)*, pages 403–417. Springer.

Sun, W., Yan, L., Ma, X., Wang, S., Ren, P., Chen, Z., Yin, D., and Ren, Z. (2023). Is chatgpt good at search? investigating large language models as re-ranking agents. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Thakur, N., Reimers, N., Rücklé, A., Srivastava, A., and Gurevych, I. (2021). Beir: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. *Proceedings of the 35th Conference on Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*.

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

Wang, S., Liu, Y., Xu, Y., Zhu, C., and Zeng, M. (2021). Want to reduce labeling cost? gpt-3 can help. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4195–4205.

Zhuang, H., Qin, Z., Jagerman, R., Hui, K., Ma, J., Lu, J., Ni, J., Wang, X., and Bendersky, M. (2023). Rankt5: Fine-tuning t5 for text ranking with ranking losses. *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*.