Evaluating Table Extraction Strategies for RAG Pipelines over Portuguese PDF Documents

Danielle Tavares da Silva[1], **Júlia Souza de Mendonça**[2], **Vitória Sofia Barbosa Pereira**[1], **Sávio Teles**[1]

[1] Instituto de Informática - Universidade Federal de Goiás (UFG)
74690-900 - Goiânia - GO - Brazil

[2]Pontifícia Universidade Católica de Goiás (PUC Goiás)
74605-010 - Goiânia - GO - Brazil

d.tavares1203@gmail.com, juliasmendonca15@gmail.com,
vitoria_sofia@discente.ufg.br, savioteles@ufg.br

Abstract. *Tabular data in PDFs challenges Retrieval-Augmented Generation (RAG) pipelines because standard extraction loses the row-column structure that gives tables their meaning. We study how table representation and retrieval affect end-to-end RAG over Portuguese occupational safety documents, crossing two extraction strategies, rule-based parsing (PDFplumber) and vision-based description by a multimodal LLM, with two retrieval strategies, dense vector search and a dense–sparse hybrid. We evaluate the resulting configurations across nine generator LLMs on a 50-pair manual benchmark and a 221-pair human-validated synthetic benchmark, scoring answer quality with a panel of five LLM judges from distinct provider families and retrieval quality against chunk-level ground truth. Contrary to the intuition that a lexical signal aids exact-match identifiers, dense retrieval outperforms the hybrid scheme across all configurations: in a lexically homogeneous procedural corpus, sparse matching cannot discriminate between candidate chunks and displaces the correct one. Retrieval is the binding constraint on performance, with unrecoverable failures concentrated on atomic operational identifiers; above this floor, generator verbosity governs outcomes, and vision-based descriptions improve quality over rule-based extraction downstream of retrieval.*

1. Introduction

Occupational safety regulation in Brazil is governed by Regulatory Norms (NRs) issued by the Ministry of Labor, complemented by internal operational procedures adopted in high-risk industries such as energy, construction, and mining. These documents are technically dense and updated frequently, making them a natural candidate for question-answering systems that let workers query their contents in natural language. Retrieval-Augmented Generation (RAG) pipelines have emerged as the dominant architecture for this kind of application, combining retrieval of relevant passages with the generation capabilities of large language models (LLMs) to produce grounded answers [Lewis et al. 2020].

A recurring obstacle in deploying RAG over real-world collections is tabular data [José et al. 2024, Sui et al. 2024], which encodes much of the critical information in occupational safety documents. Standard text extraction treats PDF content as

a linear token sequence, breaking the row-column relationships that give tables their meaning [Adhikari and Agarwal 2024]. Structured extractors such as PDFplumber preserve cell boundaries, but the serialization is still hard for LLMs to interpret when headers are hierarchical or cells span multiple lines [Sui et al. 2024]. When the answer is embedded in a table, the retrieved context is often malformed or incomplete, and the LLM produces an incorrect or unsupported answer regardless of its capability [Hoang et al. 2025, Sobhan and Haque 2025]. The open question is therefore not which LLM is most capable, but how tabular content should be represented and retrieved so that the correct information reaches the generator in a usable form.

Our objective is to determine which combination of table representation and retrieval strategy best supports tabular question answering over Portuguese occupational safety documents, along two axes. The first is the extraction strategy: rule-based structural extraction with PDFplumber versus vision-based description, in which table pages are rendered as images and interpreted by a multimodal LLM. The second is the retrieval strategy: dense vector retrieval versus a dense-sparse hybrid that adds a lexical signal, motivated by the difficulty dense retrieval faces with the exact-match identifiers common in procedural documents. We evaluate the resulting configurations across nine generator LLMs, assessing answer quality with a panel of five LLM judges from distinct provider families and retrieval quality directly against chunk-level ground truth, on a manual benchmark and a larger, rigorously human-validated synthetic benchmark.

The contributions are: (i) a systematic evaluation across nine LLMs that isolates the impact of tabular representation from generator capability over Portuguese occupational safety documents; (ii) the finding that dense retrieval outperforms a standard RRF hybrid setup, as lexical homogeneity in procedural documents causes sparse matching to displace correct tables; and (iii) evidence that structured vision-based representation insulates models from formatting noise, letting concise models match or exceed larger, verbose generators.

2. Related Work

LLM table understanding is an active area of research, as models increasingly process structured data beyond plain text. Studies evaluating how well LLMs comprehend tabular structures highlight strengths in semantic reasoning but limitations in structural awareness, numerical reasoning, and robustness across formats [Sui et al. 2024]. Other approaches add retrieval mechanisms to handle large-scale tables, processing millions of tokens over extensive tabular contexts [Chen et al. 2024].

Table extraction follows two paradigms. Rule-based tools (e.g., PDFplumber, Camelot) reconstruct tables by interpreting native PDF geometry, using ruling lines or whitespace gaps [Adhikari and Agarwal 2024]; though efficient, they degrade on complex layouts, merged cells, and multi-page tables. Vision-based approaches submit page images to multimodal LLMs such as GPT-4o or Gemini, which interpret visual structure directly. Empirical comparisons show multimodal models surpass traditional pipelines on cell-level extraction, while specialized vision models retain an advantage in layout detection [Nunes et al. 2025]. Vision-based extraction has also proven viable in Brazilian Portuguese technical settings, such as extracting property features from images for tax assessment [Ribeiro et al. 2025].

RAG over plain text is well established, but applying it to documents with tables raises specific challenges [Sobhan and Haque 2025, Hoang et al. 2025]. A central concern is transforming structured elements into semantically rich text before indexing: for tables, strategies include converting HTML tables into descriptive sentences via an auxiliary LLM and serializing Markdown tables into compact dictionaries, both making row-column relationships explicit [Sobhan and Haque 2025]; visual elements are handled through vision-language captions that render non-textual content retrievable by dense search [Hoang et al. 2025]. These findings indicate that effective RAG over structured documents depends on normalizing document diversity into a coherent textual stream.

Position of This Work. This work bridges PDF table extraction, multimodal document understanding, and RAG by evaluating extraction strategies through their downstream utility rather than extraction fidelity in isolation. Evaluating on a lexically homogeneous Portuguese procedural corpus, our findings pivot the usual RAG narrative: structured semantic representations matter, retrieval is the binding constraint, and the presumed benefit of hybrid retrieval breaks down in highly repetitive technical domains.

3. Methodology

3.1. Dataset

We evaluate on two datasets: a manually curated set of 50 question-answer pairs and a synthetically generated set of 221 pairs that underwent human-in-the-loop verification. The expert-curated manual set provides a high-fidelity probe against human reasoning, while the synthetic set provides evaluation volume.

The corpus comprises Brazilian occupational safety documents from two sources: Regulatory Norms (NRs) from the Ministry of Labor, which define the legal framework, and internal procedures from an energy-sector organization, which translate the NRs into equipment- and task-specific instructions. The internal procedures span operation, maintenance, and isolation routines, and their tables are dense with atomic operational identifiers, such as equipment codes, breaker references, and lockout-card markers, tied to specific assets. The NR tables, by contrast, encode general regulatory parameters such as exposure limits and delimitation radii. This distinction matters for retrieval: the atomic identifiers that dense vector search struggles to discriminate originate in the internal procedures, not in the public NRs.

The manual benchmark consists of 50 question-answer pairs constructed through a two-person process to ensure reliability despite its modest size. One annotator, a researcher familiar with the occupational safety domain, authored each pair by reading and interpreting the source table, identifying its key concepts and relationships, formulating a representative question a domain user might pose, and extracting the reference answer using only information contained in the table. A second researcher then independently reviewed every pair against its source table, checking that the question was answerable from the table alone, that the reference answer was correct and complete, and that no information external to the table was introduced; disagreements were resolved by revising the pair until both annotators agreed. Curating accurate queries over dense safety tables requires this manual effort, which makes large-scale annotation expensive and motivates the complementary synthetic benchmark.

The synthetic dataset was generated by a two-stage pipeline. First, tables were rendered as PNG images using PDFplumber and `pdf2image`: internal procedures, whose tables span entire pages, were captured as full pages (consecutive pages of a multi-page table stacked into one image), while the discrete tables in the NRs were cropped individually, expanding the bounding box upward to include titles and headers; candidate regions were filtered by minimum row/column counts and cell-fill ratio. Second, each image was submitted to `gemini-3.1-flash-lite-preview` via the Google GenAI vision API and prompted to produce exactly two question-answer pairs per image: natural-sounding domain queries, with no reference to tabular structure, whose answers are grounded only in the image, returned as JSON with the question, answer, and a supporting excerpt.

This yielded an initial pool of 408 pairs. Because the occupational safety domain requires factual grounding, a human annotator inspected every generated question, answer, and excerpt against its original source image to filter out hallucinations, ungrounded claims, and ambiguous phrasing. Through this process, 187 pairs (45.8%) were excluded and 10 (2.5%) corrected, yielding a verified evaluation set of 221 pairs.

3.2. Table Extraction Strategies

3.2.1. PDFplumber Extraction.

PDFplumber extracts structured content from native PDF files, recovering text, tables, and positional information. Rule-based parsers degrade on complex layouts and multi-page tables [Adhikari and Agarwal 2024]; we adopt PDFplumber as the representative rule-based extractor because it accesses character-level bounding boxes directly from the digital PDF stream, representing the ceiling for non-visual structural parsing on digital-native documents. It parses each document and extracts table regions as flat delimited text. While recent work demonstrates that serializing tables into structured formats like Markdown or HTML is crucial to preserve row-column semantics for LLMs [Sui et al. 2024], we deliberately retain the raw delimited output. This establishes a "naive" structural baseline, representative only of legacy or out-of-the-box RAG defaults lacking format normalization. We explicitly caution that this choice bounds our findings: comparing this raw geometric extraction against generative semantic descriptions isolates the impact of substituting a rudimentary mechanical parser with an LLM-reasoned representation, but does not rule out that an optimized, Markdown-serialized rule-based pipeline could close the observed performance gap.

3.2.2. Vision-Based Extraction.

We combine the page-level detection of PDFplumber with vision-based extraction by a multimodal LLM, which has been shown to outperform traditional OCR for table extraction [Nunes et al. 2025, Ribeiro et al. 2025]. PDFplumber identifies which pages contain tables; only those pages are rendered as images and sent to the multimodal model. To test whether the benefit stems from the visual architecture itself or from a particular model, we instantiate the pipeline with two multimodal models from distinct provider families, Gemini 3.0 Flash and Claude Sonnet 4.6, holding the rest of the pipeline fixed.

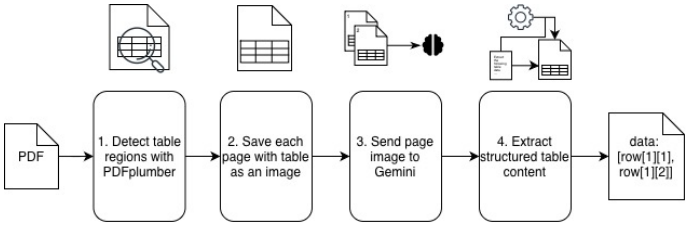


Figure 1. Pipeline for vision-based table extraction. PDFplumber detects pages with candidate table regions; each is rendered as an image and sent to a multimodal model, which returns structured JSON.

As shown in Figure 1, each flagged page is converted to an image and submitted to the model via API. The model was instructed to: (i) identify all technical tables across consecutive page images; (ii) merge multi-page tables into a single entry; (iii) ignore non-technical layouts such as headers, revision blocks, and signatures; (iv) reconstruct split cell content from lowercase continuations; and (v) return only a JSON object with the table title, first page, and row-column data. The model was called with temperature 0.0 and a maximum of 16,384 output tokens. These instructions target the two structural challenges most common in the corpus: tables that span several pages, frequent in the internal procedures, where a single procedure table continues across consecutive pages (handled by instruction ii), and cells whose content wraps onto additional lines, where a description is split across rows and resumes with a lowercase continuation (handled by instruction iv). Treating each multi-page or multi-line table as one logical unit prevents a single table from being fragmented into disconnected chunks at indexing time, which would otherwise scatter its rows across the vector space.

3.3. Retrieval Pipeline

To attribute downstream differences to representation and retrieval rather than to chunking or embedding, the indexing stage is held constant across all configurations. Each extracted table was serialized as a single chunk, retaining its title, source page, and row-column data, embedded with `gemini-embedding-001` (default 3,072 dimensions), and stored in a FAISS index. On this shared index we compare two retrieval strategies, each returning the top 5 chunks as context.

Dense vector retrieval. The first strategy isolates dense passage retrieval [Karpukhin et al. 2020]: the top 5 chunks by cosine similarity are retrieved, with no sparse matching or reranking. This homogeneous setup lets us observe how textual representations of tables interact with vector spaces, and exposes the failure modes of semantic embeddings on exact-match technical identifiers.

Hybrid retrieval. The second strategy reintroduces a lexical signal, combining the dense retriever with a sparse BM25 retriever and fusing the two rankings via Reciprocal Rank Fusion (RRF) [Cormack et al. 2009], which ranks candidates by position rather than raw score. While RRF bypasses the direct scaling issues of BM25 and cosine similarity without explicit score-weight calibration, equal-weight fusion is an uncalibrated baseline that assumes both retrieval spaces are equally informative. We use this standard, unweighted setup and truncate the fused list to the top 5.[1]

[1]Each retriever returns its top 20 candidates before fusion; BM25 uses standard parameters (k_1=1.5,

3.4. Models

Although Portuguese-specific language models exist [Souza et al. 2020, Pires et al. 2023], we focus on general-purpose multilingual models, the standard candidates for production deployment. We selected models that are openly accessible through a managed inference API and that span a range of parameter scales and provider families, deliberately including both compact (e.g., Gemma-4-26b) and large (e.g., GPT-OSS-120b) models. This spread is intentional rather than a confound: because our research question is whether the table representation or the generator governs answer quality, contrasting models across scales lets us test whether parameter count or the retrieved context is the operative factor, a question we return to in Section 4.3. We evaluated nine LLMs across six provider families: GPT-OSS-120b and GPT-5.4-mini (OpenAI); Gemini-3-Flash, Gemma-4-26b, and Gemma-4-31b (Google); DeepSeek-v4-pro (DeepSeek); Kimi-K2.5 (Moonshot AI); Qwen-3.5-27b (Alibaba); and GLM-5 (Z-AI). The six non-Google models were accessed through OpenRouter, while the three Google models were served through the Google API. All used greedy decoding (temperature 0.0) and a maximum of 4,096 output tokens, with no fine-tuning. Each model received a fixed prompt instructing it to answer using only the provided context, to reply "I did not find it in the provided context" when the answer is absent, and to answer directly without preamble.

3.5. Evaluation Metrics

Rather than a single evaluator, as is common in LLM-as-a-judge setups [Zheng et al. 2023], we assess each answer with a panel of five models from distinct provider families: DeepSeek-chat-v3.1, MiMo-v2-flash (Xiaomi), MiniMax-01, Seed-1.6-flash (ByteDance), and Qwen3-235b-a22b (Alibaba). LLM judges carry verbosity, position, and self-enhancement biases [Zheng et al. 2023]; drawing judges from independent pipelines prevents any one provider's style from determining the verdict, and a majority vote makes the outcome robust to a single dissenter. The panel is also more cost-effective than a single frontier judge, making per-question evaluation feasible across every configuration.

Each judge is queried independently at temperature 0.0 with a short output, and rates every answer on three dimensions; for the two categorical ones we also report inter-judge agreement via Fleiss' κ. *Correctness* places the answer on a four-level scale relative to the reference (identical, semantically equivalent, partially correct, or incorrect); an answer counts as correct when at least three judges assign one of the top two levels (the *correctness pass rate*). *Grounded* is a binary judgment of whether every claim is supported by the context, counted as grounded under the same majority rule (the *grounded pass rate*). *Judge Score* is a holistic 0-10 rating against the reference, where formatting- and verbosity-only differences are not penalized and an unwarranted "not found" counts as an error; it is averaged across judges and normalized to $[0, 1]$.

The validity of this automated multi-judge panel relies on our manual benchmark creation: because the reference question-answer pairs were curated and validated by hu-

b=0.75) and RRF the standard constant c=60, implemented with LangChain's `EnsembleRetriever`. We acknowledge that more sophisticated learned weights or cross-encoder reranking might mitigate the noise injected by BM25.

man domain experts, the LLM judges evaluate the generators against a controlled ground truth.

Independently of answer quality, we measure how well each extraction-retrieval combination places the correct table in the retrieved context, using chunk-level ground truth in which each question is annotated with the source document and page of its answer table. A retrieved chunk is relevant when its source and page match this annotation. We report Recall@k for $k \in \{1, 3, 5\}$, MRR, and nDCG@5, computed once per configuration since retrieval is independent of the generator.

4. Results

4.1. Extraction and Retrieval Strategies

Table 1. Answer quality averaged over the nine generators, per extraction $\times$ retrieval configuration. Best value per metric within each benchmark in bold.

		Manual (50)			Synthetic (221)		
Extraction	Retrieval	Judge	Corr.	Grnd.	Judge	Corr.	Grnd.
PDFplumber	Dense	0.828	0.767	0.938	0.647	0.537	0.820
Gemini	Dense	0.843	**0.793**	**0.953**	**0.674**	**0.573**	**0.849**
Claude	Dense	**0.846**	**0.793**	0.916	0.497	0.391	0.677
PDFplumber	Hybrid	0.755	0.684	0.864	0.584	0.477	0.756
Gemini	Hybrid	0.745	0.700	0.849	0.638	0.547	0.806
Claude	Hybrid	0.768	0.709	0.822	0.476	0.387	0.646

A consistent effect across Table 1 is that adding a sparse lexical signal degrades answer quality: every extraction strategy scores lower under hybrid than under dense retrieval, on both benchmarks and all three metrics. This contradicts the intuition that BM25 should naturally compensate for dense retrieval's known weakness with exact-match identifiers. The failure stems from the lexical homogeneity of the procedural corpus.

Decomposing the results by document type reveals that the degradation is heavily concentrated in the internal procedures: on the synthetic set, hybrid retrieval breaks 8.1% of the internal-procedure questions that dense retrieval answered correctly, compared to only 0.5% of NR questions. Because the internal procedures describe repetitive operation, maintenance, and isolation routines using a constrained vocabulary (e.g., "isolate," "verify breaker"), standard BM25 cannot effectively discriminate between candidate tables. Consequently, reciprocal rank fusion promotes lexically similar but incorrect chunks, actively displacing the correct context that dense retrieval had accurately isolated. In highly repetitive procedural regimes, sparse retrieval injects noise rather than signal. Given that dense retrieval consistently outperforms the hybrid approach in this setting, all subsequent analyses report the dense configuration.

4.2. Retrieval Quality

Retrieval performance isolates the exact point of failure for the hybrid strategy. As shown in Table 2, dense retrieval outperforms the hybrid setup on every metric across both benchmarks. This confirms at scale the lexical-noise mechanism detailed in Section 4.1: BM25 promotes structurally similar but incorrect chunks from repetitive procedures. Retrieval metrics are lower on the synthetic benchmark than on the manual one (dense Recall@5

around 0.20 versus 0.70-0.78), reflecting the predominance of internal procedures where the correct table is harder to isolate among hundreds of uniformly phrased candidates.

The performance gap between extraction methods during the retrieval phase is minimal. Dense retrieval metrics do not consistently favor one extraction technique over the others. This demonstrates that the end-to-end superiority of vision-based extraction arises downstream: the advantage lies entirely in how usable and interpretable the retrieved text is to the generator, rather than in algorithmic retrieval probability.

Table 2. Retrieval quality by extraction and retrieval strategy on both benchmarks. Best value per column within each benchmark in bold.

		Manual (50)					Synthetic (221)				
Extraction	Retrieval	R@1	R@3	R@5	MRR	nDCG	R@1	R@3	R@5	MRR	nDCG
PDFplumber	Dense	0.47	**0.62**	**0.78**	**0.589**	**0.629**	0.113	0.186	**0.244**	**0.163**	**0.183**
Gemini	Dense	**0.52**	0.58	0.70	0.574	0.604	**0.136**	0.176	0.195	0.161	0.170
Claude	Dense	0.50	0.60	0.74	0.576	0.616	0.118	**0.190**	0.204	0.156	0.168
PDFplumber	Hybrid	0.49	0.51	0.54	0.514	0.517	0.079	0.167	0.217	0.130	0.153
Gemini	Hybrid	0.42	0.50	0.52	0.461	0.476	0.122	0.165	0.190	0.149	0.158
Claude	Hybrid	0.44	0.52	0.54	0.481	0.496	0.104	0.145	0.167	0.127	0.136

4.3. Per-Model Behavior

As reported in Table 3, conditioned on successful retrieval, generator choice has minimal impact on top-performing models: seven of the nine LLMs cluster tightly at the top of the manual benchmark, their spread smaller than the variance induced by the extraction strategy. This confirms that the tabular representation format, rather than model parameter count, governs reasoning accuracy. The notable exceptions are GLM-5 and Qwen-3.5-27b, the only models whose performance shifts substantially with the extraction method: both degrade under raw PDFplumber serialization but recover competitively under vision-based natural-language descriptions.

Table 3. Per-model Judge Score under dense retrieval on the manual benchmark, by extraction strategy. Best per model in bold.

Model	PDFplumber	Gemini	Claude
DeepSeek-v4-pro	**0.879**	0.878	0.852
Gemini-3-Flash	**0.865**	0.864	0.851
Gemma-4-26b	**0.907**	0.882	0.830
Gemma-4-31b	0.878	**0.885**	0.858
GLM-5	0.638	0.719	**0.841**
GPT-5.4-mini	0.869	0.865	**0.879**
GPT-OSS-120b	0.868	**0.876**	0.864
Kimi-K2.5	**0.854**	0.832	**0.854**
Qwen-3.5-27b	0.694	0.782	**0.787**

These two models are also the most verbose in our evaluation, generating several times more tokens than their counterparts. Raw delimited tables force a verbose model to locate answers within unstructured geometric text, often leading to over-generation and reasoning drift. Conversely, a vision-based description supplies the semantic structure upfront, effectively constraining and focusing the model's output. The multi-judge panel corroborates this behavior: inter-judge agreement is high on Correctness (Fleiss' $\kappa = 0.75$ manual, 0.65 synthetic) but lower on the more subjective Grounded metric ($\kappa = 0.38$ and

0.41). Disagreement is concentrated almost entirely on the verbose outputs of GLM-5 and Qwen-3.5-27b, where judges split on whether surplus conversational tokens should be penalized. Structured vision-based extraction thus insulates even verbose models from formatting noise, helping them maintain factual precision.

4.4. Latency and Token Usage

Table 4 reports per-model latency and token usage on the manual dataset, aggregated over all configurations (300 queries per model). We report median rather than mean latency, which is inflated by serving outliers (timeouts and retries).

Table 4. Median latency and token usage per model on the manual dataset, aggregated over all configurations. Serving indicates OpenRouter vs. direct Google API.

Model	Serving	Latency (ms)	Out tok.
GPT-5.4-mini	OpenRouter	979	18
Gemini-3-Flash	Google	2888	446
Gemma-4-26b	Google	2953	66
GPT-OSS-120b	OpenRouter	3104	162
Gemma-4-31b	Google	4742	68
DeepSeek-v4-pro	OpenRouter	8484	279
Kimi-K2.5	OpenRouter	13876	520
GLM-5	OpenRouter	19414	1015
Qwen-3.5-27b	OpenRouter	22349	1448

The dominant driver of latency is output length, rather than model parameter count. Because input tokens remain nearly constant across our shared-context setup, latency scales directly with generated tokens: the most verbose models are an order of magnitude slower than the most concise ones due to sequential decoding. Verbosity is therefore uniquely penalizing in this task. The longest-output models, GLM-5 and Qwen-3.5-27b, exhibit the weakest answer quality (Table 3), meaning their extra generation buys only latency without improving accuracy.

Conversely, the Gemma-4 models pair high answer quality with low token usage: Gemma-4-31b attains the highest mean Judge Score and Gemma-4-26b follows by a negligible margin, both among the most concise in the panel. As latencies across serving routes are not strictly comparable, we ground this efficiency claim in output token usage, an intrinsic property of the model.

5. Discussion and Limitations

5.1. Discussion

Retrieval sets a floor, but generation matters above it. Separating retrieval from generation clarifies where performance is decided. On the synthetic benchmark, the correct table is retrieved within the top 5 chunks for only about a fifth of the questions (Table 2). Under our strict top-5 context constraint, these retrieval misses impose an absolute ceiling on generation performance, as the reference is completely absent from the prompt. This highlights that, at a fixed and constrained retrieval budget, performance is bounded primarily by the retriever's capability over the dense corpus rather than the reasoning power of the downstream generators. Above this retrieval floor, a clear interaction emerges

between tabular representation and model verbosity: highly verbose models experience reasoning drift when parsing raw geometric text, whereas concise architectures maximize the utility of structured descriptions. Furthermore, the same lexical homogeneity explains why hybrid retrieval degraded results: BM25 cannot discriminate between lexically similar but distinct chunks and displaces the correct one, indicating that naive hybrid retrieval is detrimental where the surrounding text is repetitive.

Relation to prior work. Our finding refines, rather than contradicts, the prevailing view that dense retrieval alone is insufficient for technical-document QA. [Sobhan and Haque 2025] address this insufficiency with a fine-tuned LLM reranker over the dense candidates, and [Hoang et al. 2025] with hierarchical semantic indexing; both add a *learned* refinement stage on top of dense retrieval. Our result is consistent with their shared premise that plain dense retrieval leaves headroom, but it identifies a specific intervention that backfires in our setting: injecting an *unlearned* lexical signal through BM25 and reciprocal rank fusion degrades retrieval in a lexically homogeneous procedural corpus, because the surface vocabulary that BM25 keys on is precisely what fails to distinguish candidate tables. The contrast suggests that the productive direction for procedural corpora is the learned refinement those works pursue, such as a reranker or calibrated fusion weights, rather than the naive lexical fusion we evaluate here.

A multi-judge panel mitigates evaluator bias. Beyond structurally limiting the known biases of LLM judges (Section 3.5), the panel's internal dissent is informative: agreement is uniformly high except on the most verbose generators, where judges split on whether surplus content should be rewarded or penalized (Section 4.3). The panel thus does not merely average noise; it concentrates uncertainty on the genuinely ambiguous cases.

5.2. Limitations

Domain and language scope. The corpus is restricted to Brazilian Portuguese occupational safety material, so the patterns may not transfer to other domains or languages; the role of lexical homogeneity, in particular, is a property of dense, repetitive procedural documents. Establishing whether the advantage of generative descriptions over geometric parsing is general requires replication on multilingual tabular benchmarks such as FinQA, TabFact, or TAT-QA.

Ecosystem bias and automated evaluation limits. Our pipeline rests on the Google ecosystem: `gemini-3.1-flash-lite-preview` generated the synthetic queries, `gemini-embedding-001` defines the vector space, and Gemini serves as both extraction and generation model. Human filtering of the synthetic queries mitigates generative contamination, but the vector space may still favor Gemini-compatible phrasing. And while the multi-judge panel limits any single model's influence (despite Qwen3-235b sharing a family with a generator), LLM-as-a-judge evaluation of safety-critical outputs lacks the human factual verification real-world guarantees require.

Design choices. Several choices bound our comparisons. The rule-based baseline uses raw delimited text rather than a structured serialization (Markdown/HTML with header normalization), which could narrow its gap to the descriptions. The hybrid uses naive dense-BM25 fusion; stronger sparse signals or a learned reranker could mitigate the lexical-homogeneity failure, so our result bounds naive lexical fusion rather than hybrid retrieval in general. We also fix a single embedding model, top-5 context, and one chunk

per table; ablations over these, and over row- or section-level chunking, are left to future work, as are the ingestion-time cost and latency of the vision-based extraction stage. Finally, each model was evaluated once per configuration, so only the broader gaps discussed above are robust to the resolution of this study.

6. Conclusion

We studied how table representation and retrieval affect end-to-end RAG over Brazilian Portuguese occupational safety documents, crossing rule-based and vision-based extraction with dense and hybrid retrieval across nine generators, evaluated by a five-judge panel and by retrieval metrics against chunk-level ground truth. Dense retrieval outperformed an unweighted hybrid scheme across all configurations: in a corpus of lexically similar procedures, a naive sparse signal failed to discriminate between candidate chunks and displaced correct contexts, indicating that uncalibrated hybrid search is detrimental in repetitive domains. Retrieval remains the primary bottleneck: a substantial share of questions fail across all generators because atomic operational identifiers are never retrieved. Above this floor, generator choice substantially affects outcomes.

While vision-based descriptions outperformed the rule-based baseline end-to-end, the gap reflects the failure of unformatted geometric parsing rather than a general superiority of vision models. Isolating the true marginal value requires comparing it against an optimized heuristic parser.

These results point to retrieval as the principal lever for improving tabular RAG over procedural collections, specifically to methods that discriminate among lexically homogeneous candidates without the noise of naive lexical fusion. Among generators, the Gemma-4 models stand out as a practical default, pairing high answer quality with concise, low-cost generation. Future work should test whether these patterns hold beyond Portuguese procedural documents, replicating on multilingual tabular benchmarks and across corpora with greater lexical diversity.

7. Acknowledgments

This work was supported by the ANEEL Research and Development (R&D) Program under Project PD-02949-0124/2024 - “SafeAI: Plataforma Inteligente de Comunicação, Monitoramento e Gestão de Segurança”, funded by Foz do Chapecó Energia and developed in partnership with the UFG, CEIA, 4VANTS and Solumar. This work was also supported by the National Institute of Science and Technology (INCT) in Responsible Artificial Intelligence for Computational Linguistics and Information Treatment and Dissemination (TILD-IAR) grant number 408490/2024-1.

References

Adhikari, N. S. and Agarwal, S. (2024). A comparative study of PDF parsing tools across diverse document categories. *arXiv preprint arXiv:2410.09871*.

Chen, S.-A., Miculicich, L., Eisenschlos, J. M., Wang, Z., Wang, Z., Chen, Y., Fujii, Y., Lin, H.-T., Lee, C.-Y., and Pfister, T. (2024). TableRAG: Million-token table understanding with language models. *Advances in Neural Information Processing Systems*, 37:74899–74921.

Cormack, G. V., Clarke, C. L., and Buettcher, S. (2009). Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 758–759.

Hoang, T. T. U., Rajendran, M., Zhang, K., Wu, Y., and Nguyen, V. A. (2025). PDF retrieval augmented question answering. *arXiv preprint arXiv:2506.18027*.

José, M. M., Cação, F. N., Ribeiro, M. F., Cheang, R. M., Pirozelli, P., and Cozman, F. G. (2024). Question answering with texts and tables through deep reinforcement learning. In *Brazilian Conference on Intelligent Systems*, pages 339–353. Springer.

Karpukhin, V., Oguz, 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., et al. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in neural information processing systems*, 33:9459–9474.

Nunes, G., Rolla, V., Pereira, D., Alves, V., Carreiro, A., and Baptista, M. (2025). Benchmarking table extraction: Multimodal LLMs vs traditional OCR. In *Proceedings of the 1st Joint Workshop on Large Language Models and Structure Modeling (XLLM 2025)*, pages 8–15.

Pires, R., Abonizio, H., Almeida, T. S., and Nogueira, R. (2023). Sabiá: Portuguese large language models. In *Brazilian Conference on Intelligent Systems*, pages 226–240. Springer.

Ribeiro, G. R., Saraiva, P. A., Rosa, L. H., Marques, E. L., Pereira, G. L., Campos, P. M., Pascoal, L. M., and de Oliveira, S. S. (2025). Vision-language models for automated property feature extraction in tax assessment: A case study. In *Encontro Nacional de Inteligência Artificial e Computacional (ENIAC)*, pages 676–687. SBC.

Sobhan, S. and Haque, M. A. (2025). LLM-assisted question-answering on technical documents using structured data-aware retrieval augmented generation. *arXiv preprint arXiv:2506.23136*.

Souza, F., Nogueira, R., and Lotufo, R. (2020). BERTimbau: pretrained BERT models for Brazilian Portuguese. In *Brazilian conference on intelligent systems*, pages 403–417. Springer.

Sui, Y., Zhou, M., Zhou, M., Han, S., and Zhang, D. (2024). Table meets LLM: Can large language models understand structured table data? a benchmark and empirical study. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, pages 645–654.

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E., et al. (2023). Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. *Advances in neural information processing systems*, 36:46595–46623.