

Structured Summaries for Retrieval-Augmented Generation in Portuguese-Language Consumer Complaints

Rafael Sant’Ana, Pedro Garcia, Luis A. Duarte
Mariana O. Silva, Adriano C. M. Pereira, Gisele L. Pappa

Computer Science Department
Universidade Federal de Minas Gerais, Belo Horizonte, Brazil
{rafael.santana, pedro.clair, luisantonio}@dcc.ufmg.br
{mariana.santos, adrianoc, glpappa}@dcc.ufmg.br

Abstract

Dense retrieval is a critical component of Retrieval-Augmented Generation (RAG) systems and is highly sensitive to the quality of document representations stored in the retrieval index. In consumer complaint settings, raw interaction texts are often lengthy and noisy, which limits retrieval effectiveness. This paper investigates whether schema-guided structured summaries can improve dense retrieval in RAG. We compare embeddings derived from raw interactions and from LLM-generated summaries in a controlled evaluation on Portuguese-language consumer complaints. Summary-based retrieval achieves a Recall@1 of 0.527, compared to 0.001 when indexing raw interactions, and reaches Recall@10 of 0.610, demonstrating gains of more than two orders of magnitude. These results show that structured summaries enable more effective and reliable retrieval at low cutoffs, making them particularly suitable for RAG pipelines.

1 Introduction

Retrieval-Augmented Generation (RAG) has become a widely adopted strategy for enhancing Large Language Models (LLMs) by grounding their responses in external knowledge sources (Gao et al., 2023; Wu et al., 2024). By combining dense retrieval with neural generation, RAG systems aim to improve factual accuracy and contextual relevance of generated outputs. In customer service and complaint handling scenarios, RAG enables organizations to retrieve relevant past cases and support decision-making processes such as negotiation, mediation, and conflict resolution (Xu et al., 2024; Lima et al., 2025).

Despite its promise, the effectiveness of RAG systems depends on the quality of the document representations stored in the retrieval index. Real-world consumer complaint data are typically long, informal, and highly redundant (Félix et al., 2018;

Rabbi et al., 2024). These characteristics can reduce the discriminative capacity of dense embeddings, hindering the retrieval stage and leading to mismatches between user queries and relevant documents. In a RAG pipeline, such retrieval failures tend to propagate downstream, amplifying their negative impact on the generation component and ultimately limiting system usability (Chen and Chen, 2008; Tran et al., 2020).

A common strategy to mitigate these issues in retrieval-based systems is to summarize interaction logs before indexing them, aiming to produce more concise and semantically focused document representations. Prior research on summarization for retrieval has explored summaries as alternative or complementary representations to reduce redundancy and noise in long texts. These studies often report improvements in precision, but mixed or negative effects on recall, as summarization may remove discriminative lexical cues that are useful for matching queries to documents (Sakai and Jones, 2001; Richardson et al., 2023; Pan et al., 2024).

In this work, we investigate whether structured summaries can improve retrieval effectiveness in a RAG system for Portuguese-language consumer complaint data. We propose an LLM-based summarization procedure that generates schema-guided summaries organized into three fields: *Description*, *Core Issue*, and *Resolution*. Rather than producing free-form abstracts, this structure aims to normalize heterogeneous interactions into focused, information-dense representations that are better aligned with dense retrieval objectives.

We compare two RAG setups: (i) embeddings derived from raw interaction texts, and (ii) embeddings derived from structured summaries. To isolate the effect of document representation, both configurations share the same sentence encoder, similarity function, indexing strategy, and evaluation protocol. In addition, we evaluate retrieval under different query formulations, enabling a fine-

grained analysis of how document structure and query design jointly influence retrieval quality.

Our experimental results show that embeddings derived from structured summaries substantially outperform those derived from raw interaction texts in dense retrieval. Summary-based representations achieve large and statistically significant gains in Recall@k, particularly at low cutoffs, while retrieval over raw interactions yields near-zero recall. We further show that these gains are stable across different values of k and that retrieval performance is strongly influenced by query formulation, with more detailed complaint-based queries consistently outperforming abstract core-issue queries. Our main contributions are as follows:

- We introduce a schema-guided summarization approach tailored to consumer complaint interactions, producing structured summaries optimized for dense retrieval in RAG systems.
- We provide a controlled evaluation of dense retrieval over raw interaction texts versus structured summaries on a Portuguese-language consumer complaint dataset.
- We analyze the interaction between document representation and query formulation, assessing how summaries and query design jointly affect retrieval effectiveness.

2 Related Work

Research on summarization for information retrieval has a long history, motivated by the need to reduce redundancy and enhance relevance in large document collections. Early studies showed that using summaries as retrieval units can improve precision by filtering out irrelevant or noisy information, particularly in long or unstructured texts (Chen and Chen, 2008; Tran et al., 2020). However, these gains often come at the expense of recall, as summarization may remove discriminative terms that are crucial for effective query–document matching (Sakai and Jones, 2001).

Within Retrieval-Augmented Generation (RAG) frameworks, summarization has been explored at different stages of the pipeline. As a preprocessing step, summarization is applied before indexing to simplify document representations, reduce redundancy, and improve retrieval efficiency in lengthy datasets (Sakai and Jones, 2001; Richardson et al., 2023). As a post-retrieval step, summarization is

used to condense retrieved passages before generation, to improve factual consistency and contextual focus in the generated outputs (Subbapurmath and Kaur, 2025; Cherubin and Paverd, 2025). These complementary lines of work reflect a growing interest in balancing informativeness, retrievability, and generation quality within RAG systems.

Summarization approaches are commonly categorized as extractive or abstractive (Chen and Chen, 2008; Shakil et al., 2024). Extractive methods select salient spans directly from the source text, whereas abstractive methods generate paraphrased summaries that capture the core meaning of the document. Extractive summarization has traditionally been considered more robust for retrieval-oriented applications, as it preserves original lexical material. However, recent advances in Large Language Models (LLMs) have significantly improved abstractive summarization, enabling more effective semantic compression while preserving discourse structure (Lins et al., 2024; Shakil et al., 2024).

More recently, several studies have explored the impact of LLM-based summarization on downstream retrieval and reasoning tasks. Empirical evidence suggests that LLM-generated summaries can improve semantic alignment between queries and retrieved documents, especially in domains characterized by noisy, redundant, or conversational language (Shakil et al., 2024; Jin et al., 2024). At the same time, prior work also reports that purely abstractive summaries may omit lexical anchors that dense retrievers rely on for vector similarity computation, potentially harming recall (Sakai and Jones, 2001; Shakil et al., 2024).

In the context of consumer complaints and customer service interactions, retrieval-based and LLM-enhanced systems are increasingly adopted to support case resolution, policy compliance, and decision-making processes (Praneeth et al., 2025; Zhao et al., 2026). For example, Patel (2025) propose a RAG-based framework for e-commerce customer support that combines text retrieval with structured knowledge sources to improve answer grounding. However, consumer complaint datasets typically consist of long, multi-turn interactions between consumers and organizations, often showing redundancy and heterogeneous discourse styles (Félix et al., 2018; Rabbi et al., 2024), which pose significant challenges for dense retrieval.

Summarization provides a natural abstraction layer for such data, compressing repetitive exchanges while preserving core complaint elements

(de Vargas Feijó and Moreira, 2018). However, existing research in this domain has primarily focused on tasks such as sentiment analysis, topic classification, and satisfaction prediction, with comparatively little attention to retrieval effectiveness across different document representations (Lima et al., 2025). In particular, few studies conduct controlled comparisons of retrieval from raw interaction texts versus summarized representations.

Overall, while summarization has been shown to benefit retrieval in specific settings, its role as a document representation strategy within RAG systems for consumer complaint data remains underexplored. The literature lacks systematic evaluations that isolate the effect of summarization on dense retrieval, especially for noisy real-world datasets and non-English languages. This work addresses these gaps by empirically comparing retrieval performance over raw and schema-guided representations of consumer complaint interactions in Portuguese, providing new insights into how structured abstraction influences retrieval effectiveness in practical RAG applications.

3 Methodology

This section describes the dataset, the anonymization and summarization procedures, the embedding generation process, and the Retrieval-Augmented Generation (RAG) configurations used to evaluate the impact of structured summaries as document representations for dense retrieval.

3.1 Dataset

The dataset used in this study consists of consumer complaint interactions collected from Procon-SP Digital,¹ a governmental platform that mediates disputes between consumers and companies in the state of São Paulo, Brazil. From the broader institutional repository, we extract all complaint records associated with a single company.²

Each record corresponds to a complaint message and a multi-turn interaction between a consumer and a company representative. Individual turns are stored as JSON objects containing a role field (e.g., *consumer*, *company*) and a text field with the corresponding message. To standardize the representation, all turns belonging to the

same complaint are concatenated into a single document, where each message is prefixed with its speaker role and separated by newline characters (e.g., `consumer: . . . \ncompany: . . . \n`).

The resulting corpus comprises 21,044 unique complaint interactions. In total, the dataset contains 7,636,978 tokens, with an average length of 362.91 tokens and a standard deviation of 364.71 tokens per interaction. No filtering based on length or content is applied, ensuring that the full variability and complexity of real-world consumer complaint discourse are preserved.

Given the sensitive nature of the data, we apply a two-stage anonymization pipeline to remove Personally Identifiable Information (PII). First, a curated set of regular expression rules is used to identify and mask explicit identifiers, including phone numbers, email addresses, Brazilian document identifiers (CPF/CNPJ), URLs, and numeric account references. Second, a Named Entity Recognition (NER) model fine-tuned for Portuguese³ is employed to detect and anonymize personal names and organization names (Souza et al., 2019).

To verify the effectiveness of the anonymization process, a random sample of interactions was manually inspected after processing. The inspection focused on typical identifiers such as personal names, phone numbers, and account references, and no residual PII was observed in the inspected sample.

3.2 Summarization

To transform raw complaint interactions into structured, retrieval-oriented representations, we employ the open-source large language model *Qwen3-8B-Instruct*.⁴ This model is selected due to its strong performance in generating structurally faithful outputs across diverse formats, as evidenced by the *StructEval* benchmark (Yang et al., 2025). Since our methodology requires strict adherence to a predefined JSON schema, structural reliability is a central criterion in model selection.

All summarization is performed locally using the Ollama framework, with the decoding temperature fixed at 0.0 to ensure deterministic outputs and minimize variability and hallucinations. To further guarantee schema compliance, we adopt a constrained decoding strategy based on a Pydantic specification that enforces a valid JSON output

¹<https://consumidor2.procon.sp.gov.br/>

²This design choice enables a controlled experimental setting in which observed retrieval differences can be attributed to document representation strategies rather than variations in corporate policies, response styles, or service domains.

³<https://huggingface.co/marquesafonso/bertimbau-large-ner-selective>

⁴<https://huggingface.co/Qwen/Qwen3-8B>

structure. Specifically, each complaint interaction is summarized into three high-level semantic fields:

- **Description (*Descrição*):** An objective summary of the complaint context and interaction history.
- **Core Issue (*Problemática*):** The specific service failure or product issue motivating the complaint.
- **Solution (*Solução*):** The proposed or final resolution of the case, as reflected in the interaction.

This schema reflects the core informational dimensions of consumer complaint discourse (i.e., context, issue, outcome) while removing conversational redundancy. By explicitly separating these dimensions, the summaries aim to provide semantically focused representations that may be more suitable for dense retrieval.

Summaries are generated using a zero-shot prompting strategy. The system prompt defines the model as a summarization assistant tasked exclusively with producing concise, factually accurate, and schema-compliant summaries. The user prompt injects the anonymized complaint text and full interaction history, which the model must compress into the three structured fields without introducing external information. The full prompt template and structured output schema are provided in Appendix A.

The generated summaries are substantially more compact than the original interactions. On average, summaries contain 274.96 tokens, corresponding to approximately 75.7% of the interaction length. This reduction removes conversational redundancy while preserving the core informational content required for retrieval. Illustrative examples of the transformation from raw interactions to structured summaries are provided in Appendix B.

Evaluation Metrics. To quantitatively assess summary quality, we employ complementary metrics targeting semantic preservation, factual consistency, and human-aligned judgment. Semantic preservation is assessed using cosine similarity between sentence embeddings of the original interactions and their summaries, using the Serafim 100M Portuguese sentence encoder,⁵ a model trained for Brazilian Portuguese (Gomes et al., 2024a). Lexical and contextual overlap is further assessed using BERTScore (F1) (Zhang et al., 2020).

⁵<https://huggingface.co/PORTULAN/serafim-100m-portuguese-pt-sentence-encoder>

Table 1: Evaluation of generated summaries.

Metric	$\bar{x} \pm SD$
Cosine Similarity	0.882 ± 0.049
BERTScore F1	0.731 ± 0.051
SummaC Max	0.815 ± 0.162
SummaC Mean	0.136 ± 0.075
LLM-as-a-Judge Faithfulness	7.945 ± 1.281
LLM-as-a-Judge Informativeness	8.652 ± 0.798

Factual consistency is evaluated using SummaC, an NLI-based metric that estimates entailment between source and summary sentences. We use the InferBR-NLI model,⁶ based on the large BERTimbau model (Souza et al., 2020) and fine-tuned on the InferBR dataset (Bencke et al., 2024). Both max and mean entailment scores are reported to account for the abstractive nature of the summaries. In particular, *SummaC Max* measures whether at least one summary sentence is fully entailed by the source, indicating that core facts are preserved, whereas *SummaC Mean* averages entailment across all sentences, reflecting how abstraction and paraphrasing may reduce average entailment even if key facts remain intact.

Finally, a subset of summaries is evaluated using an *LLM-as-a-Judge* protocol with Gemini 2.5 Flash-Lite, which assigns human-aligned scores for *Faithfulness* and *Informativeness* on a 1–10 scale. The subset was selected at random and comprises 1,000 examples, providing a representative sample for human-aligned quality assessment.

Summary Quality Assessment. As shown in Table 1, the generated summaries show strong semantic alignment with the source original interactions, with a mean cosine similarity of 0.882 (± 0.049). BERTScore (F1 = 0.731 ± 0.051) further confirms substantial lexical and contextual overlap, despite the abstractive compression applied.

Factual consistency results indicate that, while individual summary sentences are strongly entailed by source segments (SummaC Max = 0.815 ± 0.162), the lower mean entailment score reflects expected abstraction and paraphrasing effects rather than factual distortion. Human-aligned evaluation corroborates these findings, with high average scores for faithfulness (7.945 ± 1.3) and informativeness (8.652 ± 0.8). Overall, these results suggest that the proposed summaries effectively

⁶<https://huggingface.co/felipesfpaula/bertimbau-large-InferBr-NLI>

balance abstraction and fidelity, making them suitable as intermediate representations for retrieval.

3.3 RAG Setups

To evaluate the impact of summary-based representations on retrieval quality, we implement two RAG setups that differ exclusively in the document representations used to populate the vector index. Both configurations employ the *Serafim 335M Portuguese* sentence encoder,⁷ a dense embedding model optimized for information retrieval in Portuguese (Gomes et al., 2024b).

All documents—either raw interactions or structured summaries—are embedded as single units, without segmentation or sliding windows. This design choice ensures that each complaint is represented holistically and avoids introducing additional confounding factors related to chunking strategies. Empirically, the vast majority of documents fall within the encoder’s maximum input length (512 tokens), making truncation rare in practice.

Importantly, the content used to construct document embeddings is strictly separated from the information used to formulate retrieval queries. In the embedded **raw** setup, document embeddings are generated exclusively from the anonymized and concatenated interaction messages exchanged between consumers and companies, explicitly excluding the original complaint text. In the embedded **summary** setup, document embeddings are generated from the structured summaries (Section 3.2), concatenating only *Description* and *Solution* fields, while excluding the *Core Issue* field to prevent trivial lexical overlap with queries.

This design prevents trivial lexical overlap between queries and indexed documents, avoiding information leakage that could artificially inflate retrieval performance. At the same time, the *Core Issue* field is still used as an alternative query input, simulating a realistic scenario in which an agent—or an automated system—may formulate a focused query to retrieve relevant cases quickly without exposing the underlying structured details in the document index.

Embeddings are stored in *ChromaDB*, a vector database for dense similarity search. To isolate the effect of document representation, we construct two independent collections:

1. **Raw Collection:** embeddings generated from anonymized interaction texts only, explicitly excluding the original complaint.
2. **Summarized Collection:** embeddings generated from the structured summaries, concatenating only the *Description* and *Solution* fields, explicitly excluding the *Core Issue*.

Retrieval in both collections uses cosine distance as the similarity metric, consistent with the embedding model’s training objective. For each query, either the anonymized complaint text or the *Core Issue* field is used to formulate the query, and the system retrieves the top- k most similar documents from each collection. This setup enables a controlled comparison between raw and summarized representations under identical retrieval conditions, while rigorously preventing any information leakage from queries into the document embeddings.

4 Experimental Evaluation

This section presents an empirical evaluation of the impact of structured, LLM-generated summaries on retrieval performance within a Retrieval-Augmented Generation (RAG) setting. We investigate whether summary-based document representations improve dense retrieval effectiveness compared to raw interaction texts, how retrieval behavior evolves as the cutoff value k increases, and how different query formulations influence retrieval quality. Our experimental analysis is guided by the following research questions (RQs):

- RQ1.** Do embeddings derived from structured summaries improve retrieval effectiveness compared to embeddings derived from raw interaction texts?
- RQ2.** How does the use of schema-guided summaries affect retrieval performance across different values of k in Recall@ k ?
- RQ3.** How does query formulation impact retrieval quality in a dense retrieval setting?
- RQ4.** Which components of a schema-guided summary contribute most to dense retrieval effectiveness?

4.1 Experimental Setup

We evaluate two dense retrieval configurations that differ exclusively in the textual representation used for indexing documents:

⁷<https://huggingface.co/PORTULAN/serafim-335m-portuguese-pt-sentence-encoder-ir>

- **Raw Interaction Retrieval:** documents correspond to the full, anonymized interaction texts exchanged between consumers and companies, excluding the original complaint text.
- **Summary-based Retrieval:** documents correspond to structured summaries, obtained by concatenating the *Description* and *Solution* fields, while the *Core Issue* field is used exclusively to formulate queries.

Queries are formulated using either the original anonymized complaint text or the *Core Issue* field and embedded with the same sentence encoder. By keeping query content strictly separate from the document embeddings, we ensure that observed differences in retrieval performance reflect the quality of the document representations rather than information leakage.

Retrieval Evaluation Metrics. Retrieval effectiveness is measured using Recall@ k , with $k \in \{1, 3, 5, 10\}$. Recall@ k captures whether at least one relevant document appears among the top- k retrieved results, making it particularly suitable for RAG scenarios, where only a small number of relevant documents are typically passed to downstream generation models. In addition to point estimates, we analyze Recall@ k curves to characterize retrieval behavior as k increases. Statistical significance is assessed using paired Wilcoxon signed-rank tests, and 95% confidence intervals are computed via bootstrap resampling over the set of evaluation queries, reinforcing the reliability of observed differences.

4.2 Results and Analysis

We now present a quantitative analysis of dense retrieval performance under different document representations and query formulations, addressing each research question in turn.

Overall Retrieval Performance. To address **RQ1**, we compare dense retrieval effectiveness for complaint-based queries using raw interaction texts and structured summaries as document representations. The two retrieval configurations differ exclusively in the indexed document representation, while the sentence encoder, similarity function, indexing strategy, and evaluation protocol are held constant. This controlled setup isolates the effect of document representation on retrieval quality.

As indicated in Table 2, retrieval over raw interaction texts yields near-zero effectiveness across

all cutoffs. Recall@1 reaches only 0.001 (95% CI [0.001, 0.002]), suggesting that the relevant complaint is rarely retrieved at the top rank. Even when the cutoff is increased, the improvement remains negligible, with Recall@10 reaching only 0.003 (95% CI [0.003, 0.004]).

In contrast, summary-based retrieval shows a consistent improvement. Recall@1 increases to 0.527 (95% CI [0.520, 0.534]), corresponding to an improvement of over two orders of magnitude relative to raw interaction retrieval. This advantage persists at higher cutoffs, with Recall@10 reaching 0.610 (95% CI [0.603, 0.617]), demonstrating that the gains introduced by structured summaries are not restricted to the top-ranked position. Paired Wilcoxon tests confirm that these differences are statistically significant ($p < 0.001$), reinforcing the robustness of the observed effect.

Overall, these results provide a clear answer to **RQ1**: embeddings derived from structured summaries substantially improve the effectiveness of dense retrieval compared to embeddings derived from raw interaction texts. Across both strict (Recall@1) and more permissive (Recall@10) evaluation settings, summary-based representations consistently outperform raw interaction embeddings.

Impact of Structured Summaries. To address **RQ2**, we analyze how schema-guided summaries affect retrieval behavior as the cutoff value k increases in Recall@ k . While Table 2 reports point-wise Recall@ k values, Figure 1 illustrates Recall@ k curves for complaint-based queries, showing how retrieval effectiveness evolves as more documents are considered.

For summary-based retrieval, Recall@ k increases sharply from $k = 1$ to $k = 3$, rising from 0.527 to 0.578, after which performance saturates and remains stable up to $k = 10$. This early saturation is desirable in RAG pipelines, indicating that a small number of retrieved documents is usually sufficient to provide relevant context for downstream generation. A similar trend is observed for core-issue-based queries, where Recall@ k increases from 0.151 at $k = 1$ to 0.199 at $k = 3$, then saturates.

In contrast, retrieval over raw interaction texts exhibits uniformly near-zero recall across all values of k , with only marginal and practically insignificant gains as the cutoff increases. This behavior confirms that increasing retrieval depth does not compensate for the representational limitations of

Table 2: Recall@k for different query types and retrieval settings.

Query	Setting	R@1	R@3	R@5	R@10
Complaint	Raw	0.001 _[0.001,0.002]	0.002 _[0.001,0.003]	0.002 _[0.002,0.003]	0.003 _[0.003,0.004]
Complaint	Summary	0.527 _[0.520,0.534]	0.578 _[0.571,0.585]	0.593 _[0.586,0.599]	0.610 _[0.603,0.617]
Core Issue	Raw	0.002 _[0.001,0.002]	0.003 _[0.003,0.004]	0.004 _[0.003,0.005]	0.006 _[0.005,0.007]
Core Issue	Summary	0.151 _[0.147,0.156]	0.199 _[0.194,0.205]	0.221 _[0.215,0.226]	0.248 _[0.242,0.253]

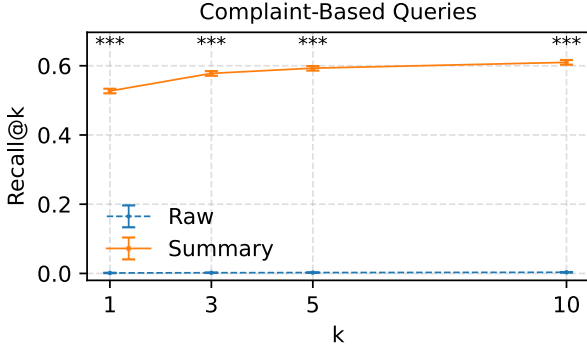


Figure 1: Recall@k curves for complaint-based queries using raw interaction texts and structured summaries. Error bars indicate 95% confidence intervals; asterisks denote statistically significant differences according to the Wilcoxon signed-rank test.

raw interaction embeddings.

Overall, answering **RQ2**, these results show that structured summaries enable effective retrieval even at low k values. In more complex scenarios, however, considering additional summary fields or incorporating re-ranking strategies could further enhance performance beyond the early saturation observed here (Gao et al., 2023).

Query Formulation Analysis. To address **RQ3**, we investigate how query formulation affects dense retrieval effectiveness when all other components of the retrieval pipeline are held constant. Specifically, we compare complaint-based queries and core-issue-based queries under the summary-based document representation.

As shown in Table 2, queries based on the full anonymized complaint consistently outperform those derived from the core issue across all evaluated cutoffs. In the summary-based retrieval setup, complaint queries achieve a Recall@1 of 0.527 (95% CI [0.52, 0.53]), whereas core-issue queries reach 0.151 (95% CI [0.15, 0.16]). This performance gap persists at higher cutoffs: Recall@10 rises to 0.610 (95% CI [0.60, 0.62]) for complaint-based queries, compared to 0.248 (95% CI [0.24, 0.25]) for core-issue queries. Paired Wilcoxon signed-rank tests confirm that these differences are

Table 3: Recall@k for different summary components.

Component	R@1	R@10
<i>Complaint-based queries</i>		
Full Summary	0.527 _[0.52,0.53]	0.610 _[0.60,0.62]
Description	0.542 _[0.54,0.55]	0.623 _[0.62,0.63]
Solution	0.006 _[0.01,0.01]	0.016 _[0.01,0.02]
<i>Core-issue-based queries</i>		
Full Summary	0.151 _[0.15,0.16]	0.248 _[0.24,0.25]
Description	0.153 _[0.15,0.16]	0.247 _[0.24,0.25]
Solution	0.047 _[0.04,0.05]	0.086 _[0.08,0.09]

statistically significant ($p < 0.001$), highlighting the advantage of richer, context-aware query formulations for retrieval effectiveness.

Overall, this analysis provides a clear answer to **RQ3**: query formulation has a substantial impact on retrieval quality in dense retrieval systems. More expressive, context-rich queries yield significantly higher recall than concise, abstract formulations, even when retrieval is performed over structured summaries. This finding underscores the importance of query design in RAG pipelines and highlights a trade-off between abstraction and retrieval effectiveness.

Contribution of Summary Components. To address **RQ4**, we evaluate the contribution of individual summary fields to dense retrieval effectiveness, considering only the fields included in the indexed embeddings: *Description* and *Solution*. Table 3 reports Recall@1 and Recall@10 (95% CI) for complaint- and core-issue-based queries.

For complaint-based queries, embeddings from the *Description* field achieve the highest recall, slightly surpassing the full summary, while embeddings from *Solution* contribute minimally. Core-issue queries show a similar pattern: *Description* dominates, and *Solution* adds little. These results indicate that descriptive content contains the most relevant lexical and contextual information, underscoring its critical role in designing schema-guided summaries for RAG pipelines.

5 Conclusion

This paper explored the impact of schema-guided summarization in dense retrieval settings for retrieval-augmented generation over large-scale consumer complaint data in Portuguese. We investigated how structured summaries affect retrieval effectiveness in comparison to raw interaction texts, how these effects evolve across different retrieval cutoffs, how query formulation interacts with summary-based document representations, and which summary components contribute most to retrieval performance.

Our results provide clear and consistent evidence that embeddings derived from structured summaries substantially outperform those derived from raw interaction texts. Retrieval from raw interactions yields near-zero recall across all evaluated cutoffs, whereas summary-based representations achieve large, statistically significant gains, particularly at low values of k . These findings demonstrate that document representation is a critical factor for effective dense retrieval in noisy, conversational contexts such as consumer complaint data.

We further demonstrated that the advantages of schema-guided summaries are stable as retrieval depth increases. Recall@ k for summary-based retrieval shows early saturation, typically by $k = 3$, indicating that relevant documents are retrieved within the top-ranked positions. This behavior is especially desirable in RAG pipelines, where only a small number of retrieved documents can be passed to downstream language models due to computational and context-length constraints.

Our analysis of query formulation revealed that retrieval quality is strongly influenced by the semantic richness of the query. Complaint-based queries consistently outperform more abstract core-issue queries, even when retrieval is performed over structured summaries, highlighting a trade-off between query abstraction and retrieval effectiveness.

Finally, by analyzing the contribution of individual summary fields, we found that the *Description* field is the main driver of retrieval performance, whereas the *Solution* field contributes minimally. This indicates that descriptive content provides the most relevant lexical and contextual cues for matching queries and should therefore be prioritized when designing schema-guided summaries for RAG systems.

Overall, these findings highlight the importance of combining schema-guided summarization with

careful query design to enable effective dense retrieval over noisy, multi-turn interaction data. Beyond the consumer complaint domain, our results suggest that structured document representations can play a key role in improving retrieval performance for other noisy, multi-turn, or semi-structured text collections.

Limitations and Future Work. Despite these encouraging results, it is important to acknowledge some limitations. First, our experiments are conducted on complaints associated with a single company, which enables a controlled evaluation but may limit the generalizability of the findings across different organizational response styles. Second, we evaluate only one sentence embedding model and do not explore alternative retrievers, such as cross-encoders or hybrid sparse–dense approaches, which may interact differently with summarized representations. Third, the summarization procedure relies on a specific LLM and a fixed schema; different models, prompts, or schema designs could affect both summary quality and retrieval performance. Finally, our evaluation focuses exclusively on the retrieval component of RAG and does not directly measure downstream generation quality or end-user impact, which are important considerations for real-world deployment. Future work will explore broader evaluation settings, investigate hybrid and re-ranking-based retrieval architectures, and study adaptive or retrieval-aware summarization strategies. In addition, extending the evaluation to end-to-end RAG scenarios, including human and automatic assessments of generated responses, represents an important direction for assessing the practical impact of structured summaries.

Acknowledgements

This work was supported by the National Institute of Science and Technology in Responsible Artificial Intelligence for Computational Linguistics, Information Treatment, and Dissemination (INCT-TILDIAR), funded by the Brazilian National Council for Scientific and Technological Development (CNPq), grant no. 408490/2024-1.

References

Luciana Bencke, Francielle Vasconcellos Pereira, Moniele Kunrath Santos, and Viviane Moreira. 2024. [InferBR: A natural language inference dataset in Portuguese](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguis-*

- tics, *Language Resources and Evaluation (LREC-COLING 2024)*, pages 9050–9060, Torino, Italy. ELRA and ICCL.
- Berlin Chen and Yi-Ting Chen. 2008. [Extractive spoken document summarization for information retrieval](#). *Pattern Recognit. Lett.*, 29(4):426–437.
- Giovanni Cherubin and Andrew Paverd. 2025. [Highlight & summarize: RAG without the jailbreaks](#). *CoRR*, abs/2508.02872.
- Diego de Vargas Feijó and Viviane Pereira Moreira. 2018. [Rulingbr: A summarization dataset for legal texts](#). In *Computational Processing of the Portuguese Language - 13th International Conference, PROPOR 2018, Canela, Brazil, September 24-26, 2018, Proceedings*, volume 11122 of *Lecture Notes in Computer Science*, pages 255–264. Springer.
- Lucas Félix, João Silveira, Washington Luiz, Diego Dias, and Leonardo Rocha. 2018. [Avaliação automática de conteúdo de aplicações de reclamação online](#). In *Anais do VI Symposium on Knowledge Discovery, Mining and Learning*, pages 49–56, Porto Alegre, RS, Brasil. SBC.
- Yunfan Gao and 1 others. 2023. [Retrieval-augmented generation for large language models: A survey](#). *CoRR*, abs/2312.10997.
- Luís Gomes, António Branco, João Silva, João Rodrigues, and Rodrigo Santos. 2024a. [Open sentence embeddings for portuguese with the serafim pt* encoders family](#). In *Progress in Artificial Intelligence - 23rd EPIA Conference on Artificial Intelligence, EPIA 2024, Viana do Castelo, Portugal, September 3-6, 2024, Proceedings, Part III*, volume 14969 of *Lecture Notes in Computer Science*, pages 267–279. Springer.
- Luís Gomes, António Branco, João Silva, João Rodrigues, and Rodrigo Santos. 2024b. [Open sentence embeddings for portuguese with the serafim pt* encoders family](#). *Springer*, pages 267–279.
- Hanlei Jin, Yang Zhang, Dan Meng, Jun Wang, and Jinghua Tan. 2024. [A comprehensive survey on process-oriented automatic text summarization with exploration of llm-based methods](#). *CoRR*, abs/2403.02901.
- David Lima, Pedro A. Santos, and João Dias. 2025. [Retrieval-augmented generation for addressing consumer complaints](#). In *Progress in Artificial Intelligence - 24th EPIA Conference on Artificial Intelligence, EPIA 2025, Faro, Portugal, October 1-3, 2025, Proceedings, Part II*, volume 16122 of *Lecture Notes in Computer Science*, pages 121–133. Springer.
- Alex Aguiar Lins, Cecília Silvestre Carvalho, Francisco das Chagas Jucá Bomfim, Daniel de Carvalho Bentes, and Vlândia Pinheiro. 2024. [CLSJR.BR - A model for abstractive summarization of legal documents in portuguese language based on contrastive learning](#). In *Proceedings of the 16th International Conference on Computational Processing of Portuguese, PROPOR 2024, Santiago de Compostela, Galicia/Spain, March 12-15, 2024, Volume 1*, pages 321–331. Association for Computational Linguistics.
- Min Pan, Teng Li, Yu Liu, Quanli Pei, Ellen Anne Huang, and Jimmy X. Huang. 2024. [A semantically enhanced text retrieval framework with abstractive summarization](#). *Comput. Intell.*, 40(1).
- Piyushkumar Patel. 2025. [Graph-enhanced retrieval-augmented question answering for e-commerce customer support](#). *CoRR*, abs/2509.14267.
- Buchepalli Praneeth, Mohana, Eshitha Chowdary Nattem, Kamala Jetti, B. K. Kavyashree, D. Rakshitha, P. Ramakanth Kumar, and K. Sreelakshmi. 2025. [Optimization of customer feedback summarization using large language models \(LLM\) and advanced retrieval-augmented generation](#). *IEEE Access*, 13:124319–124332.
- Gestefane Rabbi, Marcelo Araújo, Gabriel Kakizaki, Julia Viterbo, Julio C. S. Reis, Raquel O. Prates, and Marcos André Gonçalves. 2024. [Identificação e caracterização de reclamações duplicadas por consumidores em múltiplas plataformas](#). In *Anais do XXXIX Simpósio Brasileiro de Bancos de Dados*, pages 313–326, Porto Alegre, RS, Brasil. SBC.
- Christopher Richardson, Yao Zhang, Kellen Gillespie, Sudipta Kar, Arshdeep Singh, Zeynab Raeesy, Omar Zia Khan, and Abhinav Sethy. 2023. [Integrating summarization and retrieval for enhanced personalization via large language models](#). *CoRR*, abs/2310.20081.
- Tetsuya Sakai and Karen Sparck Jones. 2001. [Generic summaries for indexing in information retrieval](#). In *SIGIR 2001: Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, September 9-13, 2001, New Orleans, Louisiana, USA*, pages 190–198. ACM.
- Hassan Shakil, Ahmad Farooq, and Jugal Kalita. 2024. [Abstractive text summarization: State of the art, challenges, and improvements](#). *Neurocomputing*, 603:128255.
- Fábio Souza, Rodrigo Nogueira, and Roberto A. Lotufo. 2019. [Portuguese named entity recognition using BERT-CRF](#). *CoRR*, abs/1909.10649.
- Fábio Souza, Rodrigo Nogueira, and Roberto A. Lotufo. 2020. [Bertimbau: Pretrained BERT models for brazilian portuguese](#). In *Intelligent Systems - 9th Brazilian Conference, BRACIS 2020, Rio Grande, Brazil, October 20-23, 2020, Proceedings, Part I*, volume 12319 of *Lecture Notes in Computer Science*, pages 403–417. Springer.
- Nagbhushan R. Subbapurmath and Harkeerat Kaur. 2025. [RAG based product review summarization and faithfulness evaluation](#). In *Trends and Applications in Knowledge Discovery and Data Mining*

- PAKDD 2025 Workshops, ADUR, FairPC, GLFM, PM4B and RAFDA, Sydney, NSW, Australia, June 10-13, 2025, Proceedings, volume 15835 of Lecture Notes in Computer Science, pages 336–348. Springer.

Vu D. Tran, Minh Le Nguyen, Satoshi Tojo, and Ken Satoh. 2020. [Encoded summarization: summarizing documents into continuous vector space for legal case retrieval](#). *Artif. Intell. Law*, 28(4):441–467.

Shangyu Wu and 1 others. 2024. [Retrieval-augmented generation for natural language processing: A survey](#). *CoRR*, abs/2407.13193.

Zhentao Xu, Mark Jerome Cruz, Matthew Guevara, Tie Wang, Manasi Deshpande, Xiaofeng Wang, and Zheng Li. 2024. [Retrieval-augmented generation with knowledge graphs for customer service question answering](#). In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2024, Washington DC, USA, July 14-18, 2024*, pages 2905–2909. ACM.

Jialin Yang and 1 others. 2025. [Structeval: Benchmarking llms’ capabilities to generate structural outputs](#). *CoRR*, abs/2505.20139.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. [Bertscore: Evaluating text generation with BERT](#). In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Haichuan Zhao, Zisong Song, and Zhenchuan Cai. 2026. [Should AI or human agents handle customer complaints? Exploring the impact of agent type and complaint response type on recovery outcomes](#). *Journal of Business Research*, 202:115805.

A LLM Configuration

A.1 Model Selection

The summarization pipeline employs the qwen3:8b model executed through the Ollama framework. This setup enables controlled local inference and reproducible experimentation. The model is used to transform raw complaints and interaction logs into structured summaries following a predefined schema.

A.2 Structured Output Definition

To ensure consistency across generated summaries and facilitate downstream processing, the model output is constrained to a predefined JSON schema. This schema is implemented using the Pydantic validation framework, which enforces structural correctness and guarantees that all generated summaries follow the same format:

```
class ResumoReclamacao(BaseModel):
    Descrição: str = Field(description="Um resumo
    ↳ detalhado e objetivo do cenário e do
    ↳ histórico da reclamação, incluindo
    ↳ informações cruciais como datas e
    ↳ produtos.")
    Problemática: str = Field(description="O
    ↳ cerne do problema, descrevendo o que
    ↳ motivou a reclamação e a falha específica
    ↳ do serviço ou produto.")
    Solução: str = Field(description="A resolução
    ↳ proposta ou o resultado final da
    ↳ interação, se houver, focando em como o
    ↳ problema foi ou deveria ser resolvido.")
```

These fields correspond respectively to the contextual description of the complaint, the core issue raised by the consumer, and the resolution or outcome of the interaction.

A.3 Prompt Design

The summarization process follows a system–user prompt structure to guide the model toward generating structured outputs that strictly follow the predefined schema. The system prompt defines the summarization behavior and explicitly instructs the model to output only valid JSON.

System: Você é um assistente de resumo. Sua tarefa é analisar o texto do usuário e preencher o objeto JSON estritamente no formato solicitado. Seja descritivo e completo. NÃO adicione introduções ou texto fora do JSON.

User: Analise a Reclamação e Interações abaixo e gere o resumo:

Reclamação: {reclamacao}

Interações: {interacoes_autor}

Table 4: LLM hyperparameters used for structured summary generation.

Parameter	Value
Model ID	qwen3:8b
Temperature	0.0
Max Output Tokens	300
Output Format	JSON
Timeout	300s

Table 5: Examples of complaint data before and after the summarization process. Portions of the original text are truncated for readability.

ID	Complaint+Interactions (Before)	Summarization (After)
162110	<p>COMPLAINT: "vou abrir diversos chamados e se nao tiver retorno irei recorrer com advogado. ATENDER A ESTE PROTOCOLO DO PROCON [MASCARADO]/[MASCARADO]" INTERACTION: "PROCON: nan\n PROCON: Protocolo de Atendimento Atribuído. [MASCARADO]/[MASCARADO]\n CONSUMIDOR: fizeram o extorno do valor declarado!!!!!!!!!!!!!!!!!!!!\n CONSUMIDOR: Consumidor alegou que fizeram o extorno do valor declarado!!!!!!!!!!!!!!!!!!!!\n FORNECEDOR: Prezados,\n\n Segue anexa resposta ao consumidor.\n\n Atenciosamente,"</p>	<p>"Consumidor abriu diversos chamados e ameaçou recorrer com advogado se não houver retorno. Protocolo foi atribuído ao PROCON. Consumidor informou que o valor foi estornado, enquanto o fornecedor respondeu que enviou resposta ao consumidor.\n Aguarde retorno do PROCON para verificar situação do protocolo e resolver pendência com o fornecedor."</p>
165745	<p>COMPLAINT: "Comprei uma guarnição para o tanque de combustível, quando o mecânico estalou no meu carro, o mecânico falou que a borracha tá com defeito, venho perdij ajuda para o [ORGANIZACAO] para eu tá recebendo meu dinheiro de volta." INTERACTION: "PROCON: nan\n PROCON: Protocolo de Atendimento Atribuído. [MASCARADO]/[MASCARADO]\n CONSUMIDOR: Consumidor alegou que A titular da compra já fez uma reclamação no PROCON."</p>	<p>"O consumidor comprou uma guarnição para o tanque de combustível, mas ao instalar, o mecânico identificou que a borracha estava com defeito. O consumidor solicita reembolso ao [ORGANIZACAO]. O PROCON foi notificado, mas não houve resposta. O consumidor informa que a titular da compra já protocolou uma reclamação no PROCON.\n Reembolso total do valor da compra, conforme solicitação do consumidor."</p>

A.4 Execution Parameters

The model was executed using deterministic decoding settings to ensure reproducibility across runs. Table 4 summarizes the configuration used during the summarization process.

B Dataset Samples Before and After the Summarization Process

This appendix presents illustrative examples showing the transformation from the original complaint and interaction logs to the structured summary used for indexing (Table 5).