A Modular IPA Framework for Semantic Retrieval and Generation in Legal-Administrative Workflows

Arthur C. e Silva[1], Robson Costa[2], Daniel Coutinho[2], Luis Oliveira[2], Paulo Parente[2], Carlos Caminha[1,2]

[1]Programa de Pós graduação em Informática Aplicada - PPGIA - UNIFOR
Av. Washington Soares, 1321 – Edson Queiroz
60.811-905 – Fortaleza – CE – Brazil

[2]Kunumi Lab – Universidade Federal do Ceará (UFC)
Av. da Universidade, 2431
60020-180 – Fortaleza – CE – Brazil

Abstract. *The growth of document collections has made manual analysis inefficient in contexts that require synthesis and traceability. This work proposes a framework for orchestrating pipelines in the context of Intelligent Process Automation (IPA), focused on processing unstructured data. The solution integrates OCR, NLP, machine learning, and LLMs within a modular and general-purpose architecture, with support for open-source models. A case study involving more than 14,000 legal-administrative documents demonstrated that the approach enables extraction, semantic enrichment, and vector search. The results indicate gains in efficiency, organization, and support for automated document generation.*

1. Introduction

The exponential growth of document production in institutional and governmental environments has made manual analysis of large document collections increasingly difficult. In many contexts, thousands of documents must be examined to extract relevant information and build grounded reports or analyses [Almeida and Caminha 2024]. Traditional search tools typically return lists of potentially relevant documents, but do not synthesize the information, requiring the user to manually read and integrate the content to draw conclusions. When conducted manually, this process tends to be time-consuming and prone to cognitive biases, since the selection, interpretation, and prioritization of information depend on the analyst's experience, attention, and perception. On the other hand, although automatic systems can accelerate the retrieval and processing of large data volumes, many of them do not adequately preserve source traceability, making it difficult to verify the evidence used to generate certain conclusions. Thus, a gap exists between efficiency in information retrieval, the quality of the synthesis produced, and the transparency regarding the sources underlying the analyses.

Robotic Process Automation (RPA) and Intelligent Process Automation (IPA) have established themselves as relevant approaches to address this type of challenge,

[1]Demo video: `https://youtu.be/XRagAkNXFi0`
[2]Repository: `https://github.com/arthurzueiro123/Modular-IPA-Framework`

combining rule-based automation with AI cognitive capabilities, such as NLP, computer vision, and machine learning, to handle unstructured data and highly complex processes [Langmann and Turi 2022, dos S. Bezerra et al. 2025]. In this context, automated document processing plays a central role, enabling the extraction, classification, and analysis of information at scale from heterogeneous sources. The application of large language models (LLMs) has significantly expanded these capabilities: techniques such as RAG, *fine-tuning*, and *prompt* engineering allow traditional limitations in domain-specific document processing to be overcome, enabling everything from automated attribute extraction in legal and administrative texts [Berger et al. 2024] to topic classification in large document corpora [Ngai et al. 2025]. In the public sector, this potential is especially relevant for promoting transparency, accountability, and informed decision-making [Ngai et al. 2025, Juell-Skielse et al. 2022]. [Raviolo et al. 2025] illustrate this convergence with TricIA, a multi-agent system based on an open-source language model that democratizes access to complex fiscal information through conversational interfaces, operating efficiently even on modest infrastructures.

In the Brazilian context, these difficulties manifest concretely in public administration bodies responsible for document analysis and administrative decision-making. One example is the State Revenue Secretariat of Ceará (SEFAZ-CE), the body responsible for tax management, oversight, and revenue collection. Within this structure, the Tax Resources Council of Ceará (CONAT) stands out the institution where the case study of this work was conducted an administrative body dedicated to adjudicating tax proceedings and analyzing appeals submitted by taxpayers. Within this council, technical units are responsible for examining extensive sets of procedural documents, opinions, legislation, and prior decisions to support reports and votes in administrative tax proceedings. This scenario characterizes an environment marked by large document volumes and high informational complexity, in which systematic analysis, synthesis of relevant information, and traceability of the sources used become fundamental aspects for ensuring transparency, consistency, and efficiency in the decision-making process.

2. Proposed Work

This work proposes a modular framework for building and orchestrating processing pipelines for unstructured documents in the context of IPA. The solution integrates OCR, NLP, vector databases, and LLMs into independent and reusable components, decoupled from one another and combinable to form different processing flows. Processing is organized into atomic steps grouped in a layer of specialized interfaces, responsible for: text extraction via OCR, structured information extraction, vector embedding generation, LLM integration, file persistence, and text segmentation. These components are compatible with both external services and locally executed open-source solutions, such as LM Studio, llama.cpp, and Ollama. Built on this foundation, the framework provides examples of complete pipelines that demonstrate the orchestration of components in tasks such as:

- end-to-end document processing;
- RAG architectures for contextual retrieval;
- validation and interpretation of JSON structures;
- construction of reusable and adaptable flows.

The separation between atomic interfaces and pipelines promotes extensibility and architectural experimentation, allowing application in both simple document automation tasks and more complex scenarios of semantic retrieval and automated content generation.

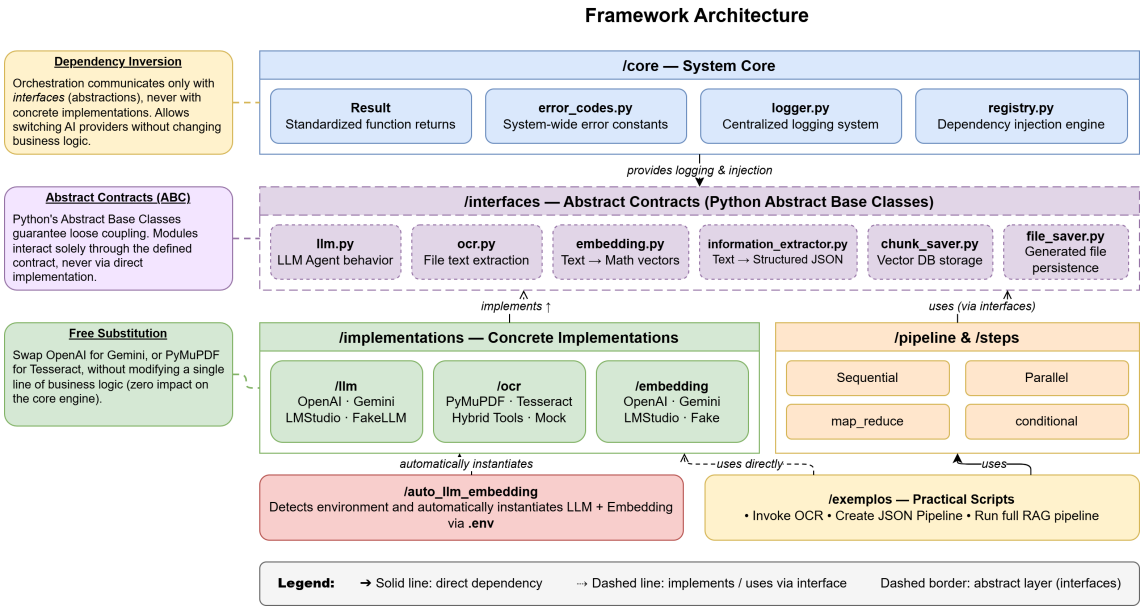

Figure 1. Framework Architecture

It is important to emphasize that, as evidenced by the proposed architecture, the goal of this work is not to compete with OCR tools, but to provide a framework for code reuse in document processing. As illustrated in the figure, any OCR implementation, such as Docling, as well as different LLMs, can be integrated into the framework, provided they implement the corresponding interface. As a result, the code remains highly decoupled, allowing existing implementations to be replaced or extended by simply changing a configuration variable. From a functional perspective, the proposed framework can be seen as analogous to a GenAI Intelligent Document Processing (GenAI IDP) platform, such as those offered by cloud providers. However, its distinguishing feature is that it is fully open source, enabling local deployment, greater flexibility, customization, and independence from proprietary cloud services.

3. Methodology

To validate the framework, a case study was conducted using real data in partnership with an extension of the Government of the State of Ceará, *SEFAZ*, operating within the CONAT Unit.

During the study, the framework was used to create a document processing pipeline focused on a real problem faced by SEFAZ, allowing its application to be evaluated in a real-world scenario.

For reasons of security, privacy, and ease of testing, the public version of the project available in the GitHub repository consists of an adaptation of the original implementation, containing some simplifications. These include storing vector database embeddings in JSON files and the default configuration for using local models via *LM*

Studio, which facilitates replicability and test execution. Furthermore, to avoid disclosing real data, a synthetic sample generated with the assistance of LLMs was made available, enabling validation of the code and the proposed pipeline.

3.1. Problem

The process of drafting the documents analyzed in this study is based on consulting a broad set of reference materials used to provide legal and administrative grounding for the decisions produced by auditors. This collection comprises more than 14,000 documents, including prior decisions and other relevant institutional records. In practice, identifying appropriate references depends significantly on the auditor's prior experience and knowledge of relevant legislation or similar cases. When the topic at hand does not fall directly within the professional's area of expertise, searching for precedents and related documents can demand considerable effort, involving the manual review of multiple files until materials capable of adequately supporting the argumentation and legal basis of the document being drafted are found.

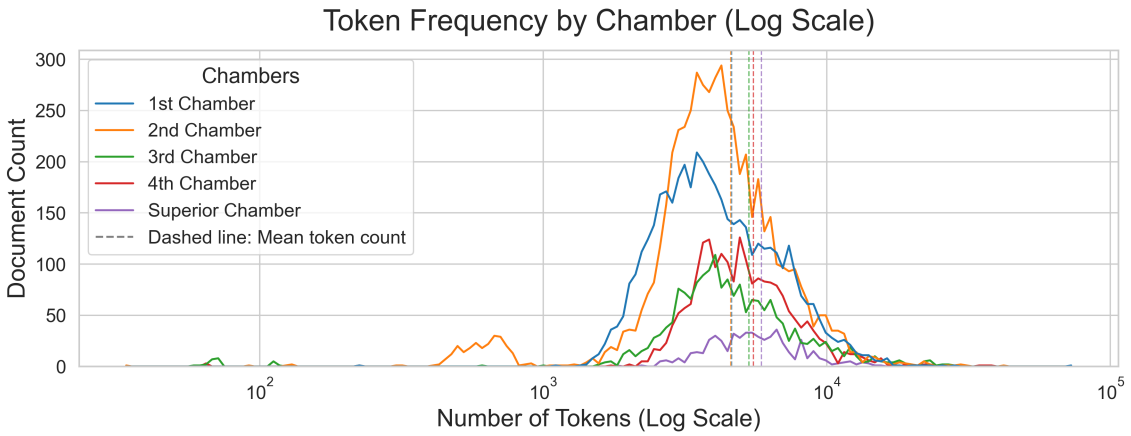

Figure 2. Token frequency distribution across documents

Folder	File Count	Total Tokens	Avg Tokens/Doc	Total Pages	Tokens/Page	Total Size (MB)
1st Chamber	4623	21130696	4570	30820	685	2023.38
2nd Chamber	5622	25534982	4541	39296	649	2830.09
3rd Chamber	1890	9844802	5208	14060	700	1309.54
4th Chamber	2146	11488895	5353	15660	733	1157.98
Superior Chamber	572	3215489	5621	4710	682	297.88

The documents comprising the analyzed collection are organized across different adjudicating chambers, which correspond to structural units responsible for reviewing proceedings. Specifically, the collection is divided among the First, Second, Third, and Fourth Chambers, as well as the Superior Chamber. The first-instance chambers (First through Fourth) are responsible for the regular analysis and adjudication of administrative proceedings, in which initial decisions are rendered. The Superior Chamber, in turn, acts as a reviewing body, responsible for examining cases in which an appeal or contestation was filed against a decision originally issued by the ordinary chambers. This organization reflects the hierarchical structure of the decision-making process and guides the classification of documents within the institutional collection.

3.2. Tool

Given the large volume of available data and the need to support the drafting of the final document, a document processing pipeline was developed aimed at analyzing and retrieving relevant information. The process begins with loading the documents that comprise the proceeding to be analyzed. Next, optical character recognition (OCR) techniques are applied, enabling the conversion of content originally in image or PDF format into processable text. From this text, relevant information is extracted from each document, forming a structured representation of the analyzed content. Subsequently, the extracted text is used as the basis for a vector search step over a previously indexed database comprising approximately 14,000 historical documents used as reference for grounding decisions. This step enables the retrieval of semantically similar documents that are potentially relevant to the case under analysis. Finally, the retrieved information and the content processed throughout the previous steps are synthesized into a single structured document. All pipeline steps make use of large language models (LLMs), with the initial development of the tool based on open-source models, including variants from the Gemma family and embedding models such as BGE, used for building and querying the vector database.

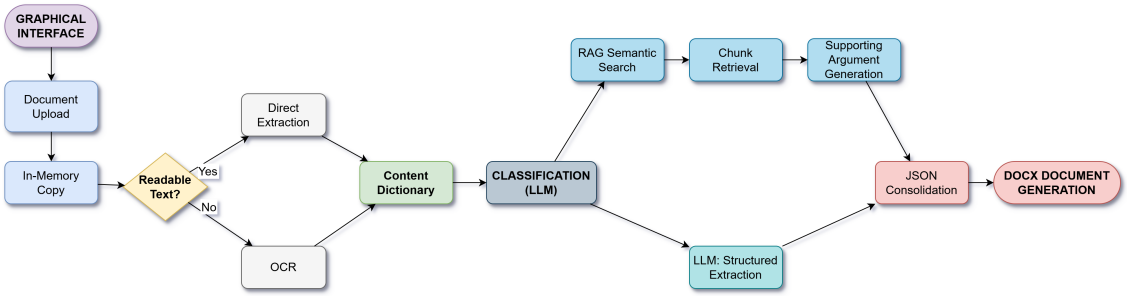

Figure 3. Flowchart of the processing pipeline implemented for the problem

The developed system features a graphical interface that allows auditors to interact with the tool in a simple and organized manner. The interface was designed to facilitate document submission, configuration of the required parameters, and automatic generation of the final document used in the analysis process.

The interface operates through the creation of records linked to administrative proceedings. Each analysis is identified by a case number, which organizes the documents and case information. The workflow spans from case entry to the generation and delivery of the final document, with support for subsequent interactions via chat.

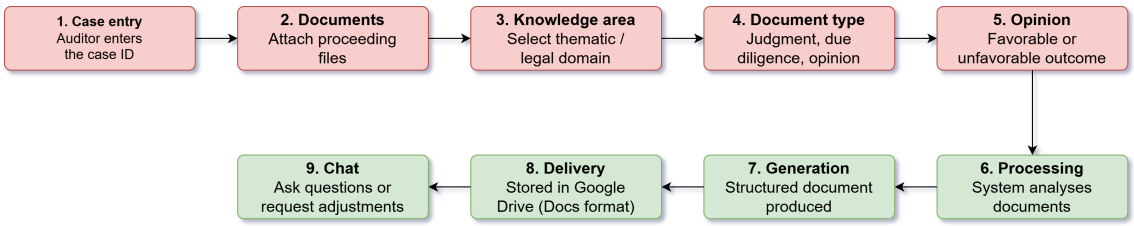

Figure 4. Flowchart of the user interaction and automated processing pipeline implemented for the case analysis framework.

4. Conclusion

The results demonstrate that the proposed framework significantly improves the efficiency of unstructured data processing. Its modular and decoupled architecture provides greater flexibility in pipeline composition while facilitating component maintenance, reuse, and evolution. In practice, improvements in both performance and organization were observed throughout the processing workflow. In the case study, the pipeline generated by the framework was used by all 14 auditors of the CONAT division at SEFAZ, which handles approximately 130 documents per month. The evaluation involved six of these auditors and showed a reduction in the average report preparation time from approximately 785 to 525 minutes. In addition, about 80% of the participants reported increased productivity and reduced cognitive effort, with no noticeable impact on the final quality of the reports. As future work, the framework will be evolved into a native Python library to facilitate its adoption, distribution, and integration into different projects. Support for additional models will also be expanded, and the orchestration and processing mechanisms further optimized, broadening the framework's applicability to real-world scenarios.

Acknowledgments

The authors thank the Kunumi Institute for supporting this research. Carlos Caminha also thanks FUNCAP (Fundação Cearense de Apoio ao Desenvolvimento Científico e Tecnológico).

References

Almeida, F. C. and Caminha, C. (2024). Evaluation of entry-level open-source large language models for information extraction from digitized documents. In *Symposium on Knowledge Discovery, Mining and Learning (KDMiLe)*, pages 25–32. SBC.

Berger, P. P., Souza, M. W., Quartezani, H., Merisio, G., Fazolo, I. A., Andrade, L. G. F., and Silva, S. T. (2024). Automação da classificação de documentos do ministério público de acordo com os objetivos de desenvolvimento sustentável utilizando processamento de linguagem natural. In *Anais Estendidos do XXXIX Simpósio Brasileiro de Bancos de Dados (SBBD)*, pages 302–307, Florianópolis, SC, Brazil. SBC.

dos S. Bezerra, A. L., Batalha, I. S., Filho, L. R. A., Almeida, C. M., Dantas, M. I. N., and Gouêa, N. A. (2025). RPAs e Data Lakes para a Indústria 4.0: Um Estudo de Caso de Ecossistema de Dados Integrados. In *Anais do XL Simpósio Brasileiro de Bancos de Dados (SBBD)*, pages 753–759, Fortaleza, CE, Brazil. SBC.

Juell-Skielse, G., Lindgren, I., and Åkesson, M. (2022). Service automation in the public sector. In *Progress in Information Systems*, pages 101–120. Springer.

Langmann, C. and Turi, D. (2022). *Robotic Process Automation (RPA): Digitization and Automation in Public and Private Sectors*. Springer, Wiesbaden, Germany.

Ngai, E. W. T., Lui, A. K. H., and Kei, B. C. W. (2025). Natural language processing in government applications: A literature review and a case analysis. *Industrial Management & Data Systems*, 125(6):2067–2104.

Raviolo, B., Chaves, J. L. M. S., Filho, R. H., and Caminha, C. (2025). Tricia: A multi-agent system to simplify interaction with the business intelligence dashboard platform of the fiscal monitor. In *Companion Proceedings of the 40th Brazilian Symposium on Databases (SBBD)*, Fortaleza, CE, Brazil. SBC.