Transforming Public Administration: AI-Driven Tools for Summarizing Corruption Cases in Federal Education

Ruan Rolim[1], **Giselle Batista**[1,2], **Pedro Bacelar**[1], **Marcela Souza**[1,2], **Hugo Kuribayashi**[1,2]

[1]ManivaLab - Grupo de Pesquisa em Transformação Digital na Amazônia Sul Oriental
Universidade Federal do Sul e Sudeste do Pará (Unifesspa) - Marabá, PA - Brasil

[2]Programa de Pós-Graduação em Ciências Forenses
Universidade Federal do Sul e Sudeste do Pará (Unifesspa) - Marabá, PA - Brasil

{ruan.vieira,gisellebatista,pedrobacelar}@unifesspa.edu.br

{marcela.alves,hugo}@unifesspa.edu.br

Abstract. *The fight against corruption in public institutions, particularly in universities and federal institutes, requires efficient tools for auditing and monitoring administrative processes. This paper presents a tool designed to analyze corruption cases by identifying key elements within these processes, such as classification, damage, and motivation, while observing recurring patterns. The tool leverages Natural Language Processing (NLP) techniques and Large Language Models (LLMs) to automate the extraction and analysis of relevant information. The results demonstrate that the solution can significantly reduce the time required for audits while enhancing accuracy in detecting irregularities.*

1. Context

The fight against corruption is widely recognized as a priority by governments, businesses, and civil society organizations on a global scale. In countries where corruption control is weak, other aspects of public administration also tend to be deficient, resulting in issues such as inefficiency in service delivery, low transparency, and generalized loss of public trust. [Topchii et al. 2021].

In this context, the 2030 Agenda, established by the United Nations, defines a set of Sustainable Development Goalss (SDGs) that serve as global guidelines to foster a more just and sustainable world [United Nations 2015]. In particular, SDG 16 highlights the importance of promoting effective, accountable, and transparent institutions, emphasizing the need to significantly reduce corruption and bribery in all their forms to strengthen governance and public trust.

The educational sector, in particular, is identified as highly vulnerable due to administrative complexity and the volume of resources involved. According to the General Comptroller of the Union - *Controladoria-Geral da União* (CGU), this area faces significant challenges related to the oversight and control of resources [Rodrigues et al. 2020]. Furthermore, these institutions are responsible not only for promoting knowledge but also for shaping civic values, which amplifies the damage caused by illicit practices.

In Federal Institutions of Higher Education – *Instituições Federais de Ensino Superior* (IFES), combating corrupt acts presents specific challenges. The lack of standardized tools for monitoring and analyzing irregularities contributes to the perpetuation

of inadequate practices. Administrative Disciplinary Processes - *Processo Administrativo Disciplinar* (PAD) are valuable sources of information for the detection and even prevention of illicit activities, as they allow for the identification of patterns that can be anticipated and mitigated in the future. However, these records face issues such as a lack of organization and a shortage of qualified human resources to perform detailed analyzes.

In general, PADs are extensive procedural documents, containing hundreds of pages that include denunciations, testimonies, evidentiary documents, and reports. This complexity renders manual analysis impractical, particularly given the limited structure of IFESs in terms of resources and personnel to address the required volume and depth. Furthermore, manual analysis of PADs is susceptible to subjectivity. Each individual involved in the evaluation process may apply different interpretation criteria, which undermines the uniformity and effectiveness of the conclusions. Additionally, the absence of automated methods for identifying patterns in corruption cases hampers learning from past processes. Learning from records of irregularities is essential to develop preventive strategies and strengthen public governance.

This landscape highlights the need for innovative solutions aimed at a scenario of digital transformation as a strategic path for continuous improvement and innovation in public service. The related literature has shown advances in the use of Artificial Intelligenve (AI) in the public sector, providing innovative solutions for audits, document processing, and irregularities detection. The use of such technologies can also help prevent corruption by identifying suspicious patterns of illegal activities before they cause significant harm [Gilson and Bramili 2023]. These studies employ advanced approaches, such as Natural Language Processing (NLP), Machine Learning (ML) and Large Language Models (LLM) [Silva et al. 2024b, Silva et al. 2024c, Silva et al. 2024a, Gilson and Bramili 2023].

Therefore, the current context underscores the urgency of initiatives that combine technology and innovation to strengthen the response capacity of IFES against corruption. Overcoming the limitations of the traditional model of manual analysis becomes a challenge to be addressed in order to promote a public administration that is more transparent, efficient, and aligned with ethical and legal principles.

2. Process

The PADs are complex procedural instruments, characterized by a large amount of documentation and a procedural rigor that requires detailed analysis and robust legal justification. Providing a concise summary is a strategy to optimize user experience. Text summarization allows for the reduction of content without compromising its essence, minimizing the reading time and the cognitive effort needed for comprehension. This approach is especially relevant in the administrative area, where PADs are situated, since the analysis of large volumes of text is a frequent and challenging task [Deroy et al. 2024, Kanapala et al. 2019].

The automatic summarization of PADs faces specific challenges due to the necessity of ensuring that essential and mandatory information, often required by regulations, is preserved in the summary. The complexity of PAD, which contains the identification of the process, the parties involved, the identified infractions, the applied penalties, and the legal foundation, requires a structured process that allows the extraction and organi-

zation of this information without compromising the integrity of the content. Although existing approaches, such as extractive and abstract summarization, have shown progress, they still face challenges related to the preservation of crucial data, which can result in disorganized or incomplete summaries [Bai and Chunglun 2024].

The proposed solution seeks to implement methods that integrate not only textual comprehension but also the structure and legal requirements of the document, to provide an effective summarization that is appropriate for the legal context of PADs. In our approach, the content of each PAD is first extracted and then divided into smaller segments (chunks) of 1,000 tokens with an overlap of 200 tokens, ensuring contextual continuity. Figure 1 presents the flow chart of the solution proposed in this work. Given a PAD, the text of its content is extracted, and using the Langchain framework, these chunks are processed with OpenIA's models: text-embedding-ada-002 converts the text into embeddings (vector representations), and gpt-4o-mini is employed for the generation of the final summarization. These embeddings allow efficient queries based on semantic similarity. Subsequently, a search for specific information is conducted based on similarity criteria, identifying content that meets predefined patterns or characteristics. This step focuses on locating chunks related to the context of the information sought.

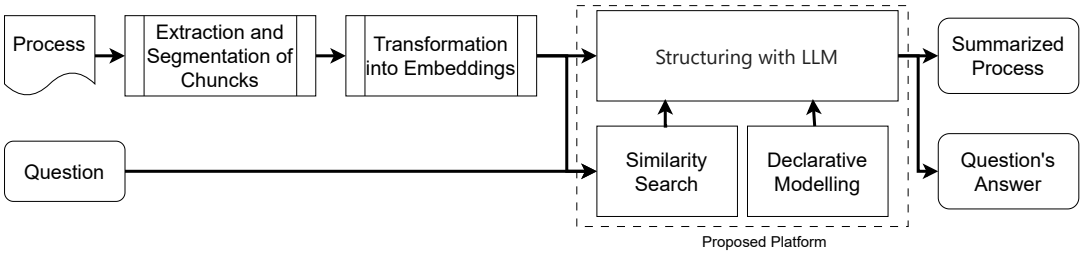

Figure 1. Representation of the information flow in the proposed tool.

Finally, based on the results obtained from the search, the proposed solution uses the LLM (gpt-4o-mini) to synthesize a structured summarization of a PAD. To ensure the validation and structuring of the summarization data, the application utilizes declarative modeling-based strategies, where each entity in the process is defined with specific attributes, well-established data types, and strict validation rules. Thus, there is a guarantee that the generated summary will be structured to present the main information of the process, such as typification, accusations, legal foundation, among other information characteristic of a PAD.

3. Solution

The proposed solution aims to contribute to innovative compliance practices in IFESs. By automating data extraction and summarization of administrative processes, the tool aims to reduce auditor time, enabling organizations to allocate human resources to strategic and preventive activities. Figure 2 illustrates the application of the tool in a real PAD, obtained through the Fala.br Platform[1].

In this context, it is important to highlight that PADs are unstructured data sources. Although each PAD has a mandatory set of information related to the intrinsic nature

[1] Available at: https://falabr.cgu.gov.br/web/home

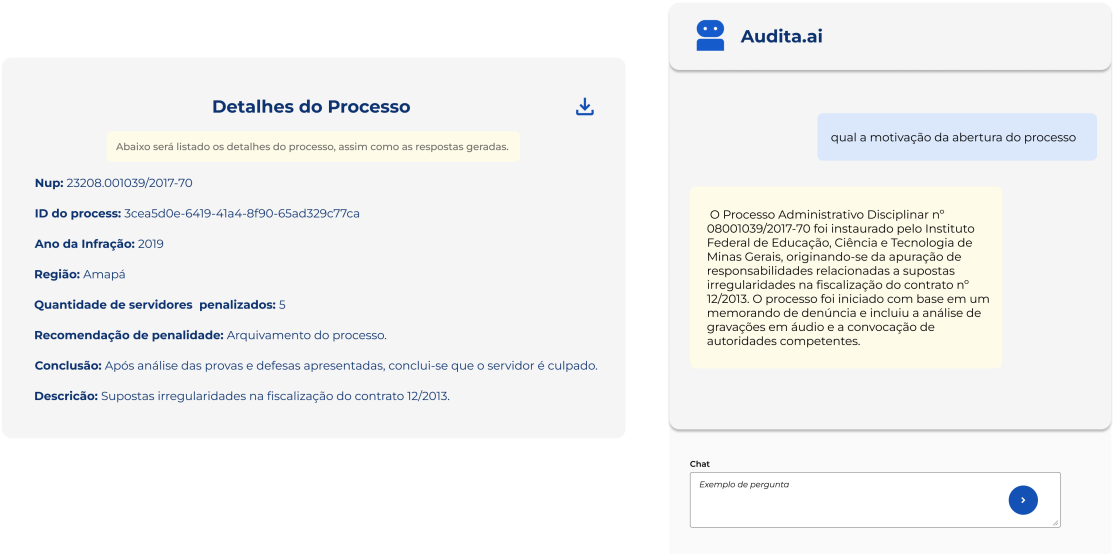

Figure 2. Representation of the summarization process for PADs, highlighting the key information extracted and organized to facilitate legal analysis.

of this type of document, these attributes should be properly extracted and incorporated into the summarization produced by the tool. This requires the application of advanced NLP techniques to ensure the precise and appropriate extraction of relevant information. Furthermore, the tool proposed in this study is capable of managing the complexity and diversity of the data contained within PADs, identifying patterns and relationships that may not be immediately apparent.

The structured integration of these data enhances the efficiency of audit processes and potentially promotes transparency and accountability within institutions, facilitating more effective and informed public management. Furthermore, this tool has the potential to significantly reduce the workload of auditors, allowing them to allocate their time and resources more strategically.

In future work, this tool could be enhanced to cluster processes with similar characteristics, allowing more in-depth analysis and the identification of recurring patterns. This evolution would contribute to greater efficiency in the detection of irregularities, making the auditing process more precise and automated. In addition, this work could lead to the establishment of a Govtech initiative focused on the development of AI-driven auditing solutions. Such a venture would streamline the auditing process across various public institutions while fostering innovation in compliance practices, creating opportunities for value generation and revenue, ultimately paving the way for a more transparent and accountable public administration.

References

Bai, H. and Chunglun, W. (2024). Mt-sal: Multi-task structure-aware learning for legal document summarization. In *2024 Cross Strait Radio Science and Wireless Technology Conference (CSRSWTC)*, pages 1–3.

Deroy, A. et al. (2024). Applicability of large language models and generative models for legal case judgement summarization. *Artificial Intelligence and Law*.

Gilson, D. H. M. I. and Bramili, G. A. (2023). Inteligência artificial no combate à fraude e corrupção: A experiência da controladoria geral do município do rio de janeiro. *Revista da CGU [online]*.

Kanapala, A., Pal, S., and Pamula, R. (2019). Text summarization from legal documents: a survey. *Artificial Intelligence Review*, 51(3):371–402.

Rodrigues, D. S., Faroni, W., Santos, N. A., Ferreira, M. A. M., and Diniz, J. A. (2020). Corrupção e má gestão nos gastos com educação: fatores socioeconômicos e políticos. *Revista de Administração Pública [online]*, 54(2):301–320.

Silva, A. L., Sampaio, V. G. R. C. A., Lima, A. M. A., Cabral, G. G., and Valença, G. (2024a). Ferramenta para auxílio à auditoria de editais municipais para compra de medicamentos. In *Anais do XX Simpósio Brasileiro de Sistemas de Informação (SBSI)*.

Silva, E. C., Medeiros, I. P., Menezes, M. V., and Kamikawachi, D. S. L. (2024b). Segmentation and summarization for extracting information about information technology equipment from government procurement notice. In *Symposium on Knowledge Discovery, Mining and Learning (KDMILE)*. Sociedade Brasileira de Computação.

Silva, M., Santos, E., Alves, K., Silva, H., Pedrosa, F., Valença, G., and Brito, K. (2024c). Using generative ai for simplifying official documents in the public accounts domain. In *Anais do Workshop de Computação Aplicada em Governo Eletrônico (WCGE)*.

Topchii, V., Zadereiko, S., Didkivska, G., Bodunova, O., and Shevchenko, D. (2021). International anti-corruption standards. *Baltic J. of Economic Studies*, 7(5):277–286.

United Nations (2015). Transforming our world: The 2030 agenda for sustainable development. Accessed: 2025-02-07.