


LLM-BASED EXTERNAL CONTROL: EMPIRICAL EVALUATION OF PRUME AI

CONTROLE EXTERNO BASEADO EM LLM: AVALIAÇÃO EMPÍRICA DO PRUME AI

CONTROL EXTERNO BASADO EN LLM: EVALUACIÓN EMPÍRICA DE PRUME AI

 <https://doi.org/10.56238/sevened2025.029-098>

Alessandro de Souza Bezerra¹, Luciane Cavalcante Lopes²

ABSTRACT

This article presents and evaluates PRUMe AI, an audit platform assisted by Language Models anchored in RAG and PROV tracks, applied to typical external control documents. Combining Design Science Research and a case study with a real sample from TCE-AM (150 documents including bids, contracts/addenda, and reports/opinions; 55% native PDFs, 45% scanned), PRUMe AI performs screening, extraction, compliance checks, and explainable reporting with structured outputs and provenance records. The results indicate material gains: average screening time from 21.4 to 7.9 min/doc (-63%) and total analysis from 39.2 to 17.8 min/doc (-55%); coverage per cycle from 25% (manual process) to 82%. In a subset annotated by experts (n=20), we obtained F1=0.86 (contract fields) and F1=0.82 (clauses), with precision@k=0.91 in the prioritization of “points of attention.” In RAG-anchored checks, 94% of findings included textual citations; the average reliability was 0.88 and inter-rater agreement reached k=0.78. PROV trails covered 96% of decisions and repetition reproduced 92% of results. We discuss limitations (OCR/layout quality, missing metadata, ambiguous wording, and curation of the normative collection) and propose an evolution agenda (document pipeline optimization, knowledge governance for RAG, and training). We conclude that PRUMe AI offers a replicable path to increasing efficiency, coverage, and standardization with transparency and auditability in external control.

Keywords: Public Audit. Artificial Intelligence. Language Models. RAG. Provenance (PROV). Explainability. Compliance.

RESUMO

Este artigo apresenta e avalia o PRUMe AI, uma plataforma de auditoria assistida por Modelos de Linguagem ancorada em RAG e trilhas PROV, aplicada a documentos típicos do controle externo. Combinando Design Science Research e estudo de caso com amostra real do TCE-AM (150 documentos entre licitações, contratos/aditivos e relatórios/pareceres; 55% PDFs nativos, 45% digitalizados), o PRUMe AI executa triagem, extração, checagens de conformidade e relato explicável com saídas estruturadas e registro de proveniência. Os resultados indicam ganhos materiais: tempo médio de triagem de 21,4 para 7,9 min/doc (-63%) e análise total de 39,2 para 17,8 min/doc (-55%); cobertura por ciclo de 25% (processo manual) para 82%. Em subconjunto anotado por especialistas (n=20), obtivemos F1=0,86

¹ Doctorate. Universidade do Estado do Amazonas. E-mail: abezerra@uea.edu.br
Orcid: <https://orcid.org/0000-0002-6410-7099>

² Master. Tribunal de Contas do Estado do Amazonas. E-mail: luciane.lopes@tce.am.gov.br
Orcid: <https://orcid.org/0009-0004-9449-5661>

(campos contratuais) e $F1=0,82$ (cláusulas), com $\text{precision}@k=0,91$ na priorização de “pontos de atenção”. Nas verificações ancoradas em RAG, 94% dos achados trouxeram citação textual; a fidedignidade média foi 0,88 e o acordo inter avaliadores atingiu $k=0,78$. As trilhas PROV cobriram 96% das decisões e a repetição reproduziu 92% dos resultados. Discutimos limitações (qualidade de OCR/layout, metadados ausentes, redações ambíguas e curadoria do acervo normativo) e propomos agenda de evolução (otimização do pipeline documental, governança de conhecimento para RAG e capacitação). Concluimos que o PRUMe AI oferece um caminho replicável para ampliar eficiência, cobertura e padronização com transparência e auditabilidade no controle externo.

Palavras-chave: Auditoria Pública. Inteligência Artificial. Modelos de Linguagem. RAG. Proveniência (PROV). Explicabilidade. Conformidade.

RESUMEN

Este artículo presenta y evalúa PRUMe AI, una plataforma de auditoría asistida por modelos de lenguaje basada en RAG y rastros PROV, aplicada a documentos típicos del control externo. Combinando la investigación en ciencias del diseño y un estudio de caso con una muestra real del TCE-AM (150 documentos entre licitaciones, contratos/adendas e informes/dictámenes; 55 % PDF nativos, 45 % digitalizados), PRUMe AI realiza la clasificación, extracción, comprobación de conformidad y elaboración de informes explicables con salidas estructuradas y registro de procedencia. Los resultados indican ganancias materiales: tiempo medio de clasificación de 21,4 a 7,9 min/doc (–63 %) y análisis total de 39,2 a 17,8 min/doc (–55 %); cobertura por ciclo del 25 % (proceso manual) al 82 %. En el subconjunto anotado por especialistas ($n = 20$), obtuvimos $F1 = 0,86$ (campos contractuales) y $F1 = 0,82$ (cláusulas), con $\text{precision}@k = 0,91$ en la priorización de «puntos de atención». En las verificaciones basadas en RAG, el 94 % de los hallazgos incluyeron citas textuales; la fiabilidad media fue de 0,88 y el acuerdo entre evaluadores alcanzó $k=0,78$. Las pistas PROV cubrieron el 96 % de las decisiones y la repetición reprodujo el 92 % de los resultados. Discutimos las limitaciones (calidad del OCR/diseño, metadatos ausentes, redacciones ambiguas y curaduría de la colección normativa) y proponemos una agenda de evolución (optimización del flujo de trabajo documental, gobernanza del conocimiento para RAG y capacitación). Concluimos que PRUMe AI ofrece una vía replicable para ampliar la eficiencia, la cobertura y la estandarización con transparencia y auditabilidad en el control externo.

Palabras clave: Auditoría Pública. Inteligencia Artificial. Modelos de Lenguaje. RAG. Procedencia (PROV). Explicabilidad. Conformidad.

1 INTRODUCTION

Audit and public control activities face a scenario of continuous transformation driven by the digitalization of processes and the massification of administrative data. Contracts, invoices, management reports and official communications are generated in increasing volumes and widely diverse formats, which require methods capable of dealing with heterogeneity and scale. Although the advancement of the public sector digitization process has increased the availability of evidence, the increase in informational complexity and diversity tends to outpace the capacity for manual analysis, resulting in evaluation bottlenecks, coverage asymmetries, and variations in the quality of findings [1].

The imbalance between data supply and analytics creates gaps in time and scale. Over time, late validation reduces the timeliness of the control and limits the possibility of correction before damage consolidation. On the scale, the definition of strongly restricted samples imposes risks of non-detection of irregularity patterns that manifest themselves in a sparse but recurrent way. In addition, the fragmentation of audit trails and the lack of standardization in the documentation of findings hinder the reproducibility of analyses and accountability [2].

The Courts of Auditors in Brazil have been adopting innovative solutions based on Artificial Intelligence (AI) to give scale and timeliness to external control, with recent evidence of diffusion and institutional maturation. A survey by IRB/Atricon indicates that, in 2024, the use of AI in External Control grew from 18 to 28 courts, also moving training and internal governance; at the federal level, the OECD classified the TCU as an advanced use case of generative AI in the public sector, highlighting the institutional offer of the technology to civil servants [3].

From a practical perspective, initiatives aimed directly at the inspection of public notices and processes are emerging. The TCE-SC developed the VigIA system, which analyzes bids before publication and flags inconsistencies for concomitant action: between 4/18 and 10/8 (2024), 7,711 notices were processed, generating 63,445 automatic responses and resulting in 215 corrections (revocations/rectifications, etc.), with materiality in the order of R\$ 2 billion [4]. Also in Santa Catarina, public notes and sectoral articles reinforce the use of VigIA for school transportation and other contracts, expanding the selection of risks to be observed and deepened by auditors [5].

Other courts have been structuring generative platforms to support document analysis and service to jurisdictions. The TCE-PE launched the Aurora platform (2024) and, in 2025, announced the evolution of AuroraChat with new functions for summarization and extraction of procedural information; the TCE-PR operates AVIA (virtual service by AI) and ChatTCEPR (institutional chat with LLM), both focused on internal productivity and service to the inspected; the TCU makes ChatTCU available and has published a guide for the responsible use of generative AI to guide safeguards, transparency, and compliance [6] [7].

In view of the context presented, we propose PRUMe AI, a web platform for control and assisted auditing that combines Natural Language Processing techniques and explainability mechanisms to support control teams in the systematic examination of large volumes of documents. The platform produces auditable trails and structured reports, prioritizing transparency, reproducibility, and adherence to applicable regulations. Unlike purely manual or black-box approaches, PRUMe AI prioritizes the interpretability of findings and the traceability of evidence.

The general objective of this study is to present the technical and functional architecture of PRUMe AI, as well as the practical results that evidence its ability to (i) increase the efficiency of the audit activity, by reducing the time of screening and analysis; (ii) expand coverage, enabling the examination of more extensive documentary universes; and (iii) strengthen compliance through systematic checking of regulatory requirements and standardization of reports.

2 THEORETICAL FOUNDATION

2.1 EXTERNAL CONTROL IN THE DIGITAL AGE

The Courts of Auditors operate in a data-intensive informational ecosystem, in which the effectiveness of control depends on the institutional capacity to transform large masses of administrative records into auditable evidence with quality and timeliness. Recent literature on public administration and digital government points out that the adoption of AI and analytics reconfigures work routines, creates new monitoring capabilities, and imposes transparency safeguards — a move that requires data governance, competencies, and standards for evidence-based decisions [8].

In this context, the use of data analytics in the public sector is growing, with empirical reports of use in control and direct administration bodies, indicating that the expansion of "big

data" changes inspection practices and service design, but also exposes challenges in data quality, interoperability, and measurement of public value [9].

For external control, the practical result is the imminent need to incorporate data science and analytics into audit cycles, with continued training and work standards that enable large-scale analysis and risk-based review.

2.2 DATA-DRIVEN AUDITING AND CONTINUOUS AUDITING

Data-driven auditing consolidates the use of analytics techniques throughout the audit cycle (planning, execution, and presentation of results), allowing both testing on complete populations and appropriate statistical sampling. In normative terms, sampling remains relevant to obtain sufficient and appropriate evidence; However, the availability of data and automation increases the feasibility of exhaustive analysis in critical areas [10].

The literature on continuous auditing describes the migration from periodic procedures to automated routines close to the event, with alerts and operational metrics integrated into the processes—a paradigm that reduces time bottlenecks and favors early detection. Case studies and academic syntheses detail principles, benefits, and implementation challenges, including the need for process reengineering and automation as a foundation [11].

For the public sector, specific guidelines from the INTOSAI community guide the conduct of audit activities with data analytics, covering concept, process, competencies and work mechanisms, and reinforcing integration with risk assessment [12].

2.3 NATURAL LANGUAGE PROCESSING (NLP) IN PUBLIC DOCUMENTS

In scenarios related to bids, contracts, reports, and other administrative and official documents, NLP enables document classification, entity and clause extraction, semantic matching, and summarization tasks, composing a typical pipeline of ingestion, OCR, extraction/normalization, and indexing/search. Recent research in the legal scope describes the challenges of legal text (length, technical language, scarcity of open data) and maps applicable tasks and models [13].

OCR quality and the characteristic noise of scans directly affect task performance *downstream*, such as NER and classification, requiring pre-processing and curation care to preserve evidence traceability [14]. On an ongoing basis, document pipelines in other regulated scopes reinforce the OCR sequence and text mining as the technical basis for mining digitized records at scale, an experience fully applicable to the public sector [15].



2.4 GENERATIVE AI AND LANGUAGE MODELS IN THE PUBLIC SECTOR

Generative AI, supported by large Language Models (LLMs), has been incorporated into government routines for assisted writing, document analysis, citizen service, and decision-making support. Recent empirical evidence shows widespread adoption in the public sector: a survey of public service professionals in the United Kingdom points to widespread use of generative tools and perception of productivity gains and reduction of bureaucratic burden, but also gaps in institutional guidelines for their use. These findings converge with the specialized literature that sees technology as an infrastructure to support services and back-office, requiring governance arrangements and specific competencies [16].

In the organizational context, the adoption of AI in the public sector involves **elements of** efficiency, fairness and transparency, although it raises challenges **implementation** related to centralization and experimentation of solutions. From a technical-methodological perspective, LLMs expand the range of applications in "Smart Government", but impose challenges of veracity, biases and security. The literature of LLMs applied to the public sector emphasizes the design of pipelines with *Retrieval-Augmented Generation (RAG)* to support responses in official documents, reduce hallucinations and enable auditability of outputs, while discussing the balance between performance and interpretability. Risk and public policy analyses on LLMs highlight regulatory implications—copyright, privacy, and algorithmic transparency—and recommend metrics and verification trails appropriate to regulated scopes [17].

The governance agenda for generative AI in public administration has advanced in an attempt to keep up with the speed of technology adoption. Recent studies systematize the risks of hallucination, *jailbreaking*, data leakage and manipulation, adopting mitigation mechanisms such as impact assessment, *guardrails* and audits. Such initiatives point to the need for regulatory capacity and accountability in the use of LLMs [18]. In courts of auditors, controllerships, and regulators, this translates into the combination of internal policies, technical compliance (provenance records, logs, access restrictions), and human review processes to preserve the legitimacy of decisions.

Finally, the literature indicates that capturing public value with generative AI relies less on isolated "proofs of concept" and more on institutional capacity directed at quality government data, redesigned processes for auditing and measuring results, and organizational learning mechanisms. Studies on the strategic adoption of AI in governments

reinforce that the critical path goes through data governance, competence management, and integration with public service performance metrics, conditions that, when met, allow LLMs and RAGs to operate as an infrastructure for efficiency, coverage, and standardization in management and control routines [19].

2.5 EXPLAINABILITY, AUDITABLE TRAILS, AND COMPLIANCE

Responsible adoption of AI in auditing requires **Explainability** (XAI) to sustain trust, review capacity and accountability. Previous research systematizes methods of explanation and discusses the balance between interpretability and performance. Additionally, techniques such as **LIME** illustrate approaches that enable agile and easy adjustments to justify predictions in text tasks [20] [21].

In terms of **Auditable trails** and reproducibility, open standards of **provenance**, such as the model **W3C PROV-DM**, allow the registration of entities, activities and agents involved in the generation of artifacts (data, models, reports), strengthening the chain of evidence maintenance. In information systems, control groups of **Auditing and accountability** guide log record generation, protection, and review as a technical safeguard [22].

As for the **Data protection** the **LGPD (Law No. 13,709/2018)** establishes principles and legal bases for the processing of data by public agencies, demanding the design of solutions with **Data Protection and Privacy Designs** through minimization, retention and security, starting from the initial phase of the information life cycle. Classic guides **Privacy Models** offer operational principles compatible with such requirements [23].

3 METHOD

3.1 ARCHITECTURE OF THE AI PLUMB

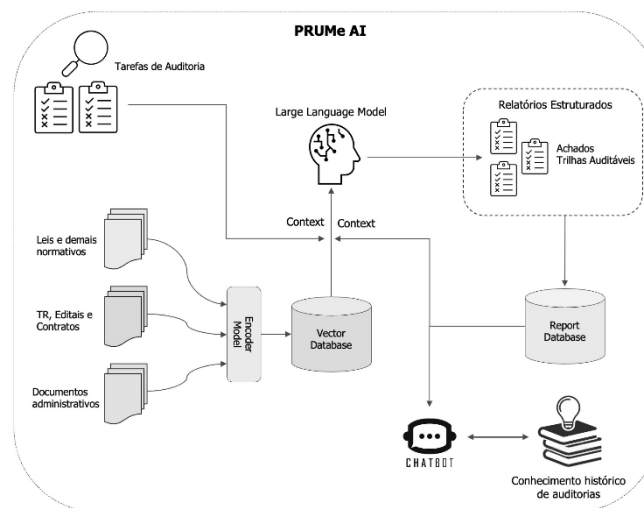
Figure 1 presents the general architecture of PRUMe AI, which can be presented through 3 main functional blocks: (i) Entry and treatment of audit data; (ii) LLM-based audit data analysis; and (iii) Presentation of results and consultation of historical knowledge of audits.

Initially, audit documents – terms of reference, legislation, general rules, notices, contracts and administrative documents – are uploaded to the PRUMe AI platform to serve as a context for analysis by the LLM. Additionally, prompts oriented to the process of analysis and identification of audit findings are submitted as a contextual complement.

The data analysis process is performed at the audit data analysis layer by LLM, processing all submitted documents under the guidelines indicated in the context prompts. Finally, in the functional layer of presentation of results and consultation of historical knowledge of audits, the results generated by PRUMe AI are presented, in addition to the availability of a Chat mechanism capable of performing queries and generating answers based on historical knowledge of audits already processed by PRUMe AI.

Figure 1

General Architecture of PRUMe AI



Source: Author

3.2 RESEARCH APPROACH

In general, we adopt a modeling process centered on artifact construction and evaluation, combining Design Science Research (DSR) for the engineering of the platform and a case study for observation and evaluation in real use. DSR guided the iterative problem cycle through the stages of: requirements, design/artifact, demonstration, evaluation and refinement, ensuring alignment between audit needs and PRUMe AI technical and architectural decisions [24].

The real-world evaluation of PRUMe AI took place through a specific case study, with multiple units of analysis (notices, contracts, and reports), explicit collection and triangulation protocol (system logs, documentary samples, and auditors' judgment), following good design and reporting practices in software engineering [25].

3.3 CORPUS AND PREPARATION

The corpus (data/documents used in the PRUMe AI analysis and evaluation process) includes notices, terms of reference, contracts/amendments, and reports in native and scanned PDF. Scanned documents undergo OCR (deskew, cleaning, recognition) and all items are subjected to layout analysis (page/blocks/tables) to preserve their semantic structures. As a technical basis and good practices, the Tesseract architecture (ICDAR'07), the PubLayNet dataset (ICDAR'19) and the LayoutLM model (KDD'20) were used [27].

3.4 LLM-ASSISTED AUDIT ARCHITECTURE

All the logic of screening, extraction, classification, normative verification, generation of findings, and elaboration of auditable trails is carried out through LLM GPT-4, accessed via OpenAI's API. The choice is based on three determining technical pillars: (a) the ability of Transformers to model long-range dependencies; (b) the performance of large-scale models for few-shot tasks; and (c) the alignment by human feedback characteristic of Instruct/GPT families, which favors useful and controllable outputs [28]. Towards **mitigate hallucinations** and ensure justification, PRUMe AI uses the paradigm of **Retrieval-Augmented Generation (RAG)**: before each decision/finding, the system consults internal databases of laws, notices and standards and **provides LLM with recovered snippets** as an obligatory context of response. The prompt passed to the LLM requires **Quotes from the recovered excerpts** whenever a normative conclusion is presented and a response is generated [29].

3.5 STRUCTURED OUTPUTS AND AUDITABLE TRAILS

LLM responses are issued in JSON with decision fields, cited excerpts, normative references, and IDs/positions in the document. Each step (OCR, retrieval, call to LLM, human validation) is recorded according to the W3C PROV: Entities (document, excerpt, find), Activities (ocr, retrieve, llm.check, report) and Agents (LLM service, auditor). The use of PROV-DM allows to reconstruct the lineage of each finding (who/what/when/how) and supports independent verification. Section 3.7 of this work provides more details about the W3C PROV. Additionally, the availability of the responses in JSON format allows us to later integrate the results into various platforms and mechanisms that simply require the standardized reading of the contents of interest [30].

3.6 TASKS PERFORMED BY THE LLM

We can organize PRUMe AI's use of LLM GPT-4 into a set of chained tasks, each of which has clearly specified inputs, procedures, and outputs. All decisions are based on prior evidence retrieval (RAG) and recorded in PROV trails, with human review at the points of greatest materiality.

I. Triage

The LLM automatically classifies the type of document to be analyzed (notice, term of reference, contract, addendum, opinion) and estimates the priority of analysis based on explicit risk rubrics (restriction of competition, object-scope misalignment, gaps in essential clauses, etc.). As data inputs, full-text elements or detected sections and basic metadata are passed. The outputs are returned in the form of type labels, risk score, and list of points of attention with reference to the relevant passages. This step reduces the initial bottleneck and directs human effort to the items with the greatest materiality potential.

II. Extraction

For classified documents, the LLM performs schema-driven extraction (JSON format) containing parts, object, values, term, supply/service items, and target clauses (readjustment, penalties, termination, guarantees, inspection). The entries focus on the segmented textual passages and, where applicable, regions of tables that can be interpreted by the OCR engine. The outputs are in the form of normalized dictionaries with IDs and positions in the document, where each field brings the source passage that supports the extraction. This structuring allows for comparable analyses between processes and facilitates subsequent verification.

III. Compliance

The compliance check is carried out by the LLM itself, but with mandatory RAG. Before responding, the system retrieves excerpts from relevant rules, drafts, and internal acts and adds them to the prompt's context. The model then confronts the extracted content with the applicable requirements, such as: the existence of an adjustment index, sanction clauses, minimum deadlines. In the outputs for each requirement, the decision (meets/does not meet/indeterminate), textual justification and citation of the normative excerpt used are defined. By requiring citation of references, the step makes the decision verifiable and auditable.

IV. Report

When there are non-conformities or weaknesses, the LLM generates structured findings containing the type (e.g., restriction to competitiveness, temporal inconsistency),

summary description, degree of severity, legal basis (with reference to the standards retrieved via RAG), and suggested correction when applicable. In addition to the text, the system produces local explanations reporting why that passage supports the conclusion, and links each justification to the evidence cited in the previous step. The goal is to standardize writing, preserve traceability, and facilitate peer review.

V. Consolidation

Finally, PRUMe IA consolidates the results in standardized reports, aggregating extracts from the documents, summary tables, and the list of findings with their respective citations. In parallel, the complete provenance is recorded in the W3C PROV standard: documents and excerpts as Entities, processing steps (OCR, retrieval, checks, generation) as Activities, and services/subjects involved (LLM, auditor reviewer) as Agents. This track allows its reexecution, independent verification and quality control over the decision-making cycle.

Figure 2 below presents a partial view of the consolidated report with the results of an analysis carried out by PRUMe AI, from the perspective of the tasks performed by the LLM and described in this section.

Figure 2

Partial view of the consolidated report of PRUMe AI

PRUMe AI — Relatório de Saída Consolidado (Execução 19/08/25)
<p>ACHADOS ESTRUTURADOS</p> <p>A seguir, são apresentados os achados gerados pelo PRUMe AI, com descrição sintética e representação JSON estruturada para auditoria e reexecução.</p> <p>A1 — Inconsistência de escopo (edital vs. TR)</p> <p>Discrepância entre objeto do edital e escopo do termo de referência, com potencial impacto na competitividade.</p> <pre>{ "id": "A1", "tipo": "Inconsistência de escopo (edital vs. TR)", "gravidade": "média", "evidencias": [{"doc": "EDITAL-2024-045.pdf", "trecho": "Objeto: contratação de empresa para manutenção predial...", "pag": 3}, {"doc": "EDITAL-2024-045-TR.pdf", "trecho": "Escopo inclui fornecimento de insumos e materiais...", "pag": 11}], "base_legal": ["Manual de minutas, Seção 2.1", "Norma interna 123/2021, art. 5º"], "acao_recomendada": "Alinhar escopo entre edital e TR, justificando materiais" }</pre> <p>A2 — Exigência potencialmente restritiva de marca</p> <p>Minuta menciona modelo/marca específica sem 'ou equivalente'.</p> <pre>{ "id": "A2", "tipo": "Exigência potencialmente restritiva de marca", "gravidade": "alta", "evidencias": [{"doc": "EDITAL-2024-071.pdf", "trecho": "Equipamento modelo X-Brand 5000", "pag": 6}], "base_legal": ["Guia de padronização de editais, item 4.3"], "acao_recomendada": "Substituir por especificação técnica com 'ou equivalente'" }</pre>

Source: Author

3.7 AI PLUMB METRICS AND EVALUATION

The verification of the quality of document extraction and identification was carried out through the observation of the metrics of accuracy, recall and F1-score in classification/extraction tasks (fields and clauses).

In binary classification problems, predictions can have four possible classes [31], they are:

True positive (PV): when the method says that the class is positive and, when checking the answer, it is seen that the class was really positive;

True negative (VN): when the method says that the class is negative and, when checking the answer, it is seen that the class was really negative;

False positive (FP): when the method says that the class is positive, but when checking the answer, it is seen that the class was negative;

False negative (FN): when the method says that the class is negative, but when checking the answer, it is seen that the class was positive.

Their equations are presented below:

- **Precision** = , which evaluates the number of true positives over the sum of all positive values, $\frac{VP}{VP+FP}$
- **Recall** = , which evaluates the method's ability to successfully detect results classified as positive, $\frac{VP}{VP+FN}$
- **F1** = $2 * \frac{Precision * Recall}{Precision + Recall}$, harmonic mean calculated based on precision and recall.

To assess prioritization, we use *precision@k* (fraction of true items among the first k alerts). In unbalanced sets, we complement it with an area under the Precision-Recall curve (PR-AUC), which is more informative than ROC-AUC [32][33].

Finally, the evaluation of explainability is carried out through the observation of the reliability/faithfulness that represents the consistency of the LLM explanation with the cited evidence (via RAG). We also conduct human judgment (double-blind) and, as a complement, LLM-as-a-Judge (G-Eval) with explicit rubrics — always with human validation to mitigate automatic evaluator bias [34].

3.7.1 Operational and Process Metrics

As a way to evaluate operational and process issues related to the use of PRUMe AI, we look at the following metrics:

- **Efficiency:** average time per document (screening + analysis) and time-to-alert (TTA) until the issuance of the finding;
- **Coverage:** proportion of the universe processed per cycle.
- **Compliance:** ratio of findings with valid normative citation (verified) on the total number of findings.
- **Standardization:** completeness of fields in the reports and consistency of structure between reports (fill/variance indexes).
- **Inter-evaluator agreement** (double validation): **Cohen's k** to measure inter-auditor reliability in labeling and citation verification.

To measure the **Inter-Evaluators Reliability** In the context of this study, we adopted a procedure of **Double validation** Led by **two auditors from the TCE-AM**, which analysed in a way **independent and blind** the **Previously described sample** (see Section 4.1). Prior to labeling, auditors went through **instruction** of use **coding** with operational definitions and examples of each category. Agreement was estimated by the ***k of Cohen***, suitable for nominal data with two judges and that **Discounts the deal at random**, accompanied by **95% confidence interval [35]**.

3.7.2 Reproducibility and tracks

As an essential element of feasibility of reproducing actions in the PRUMe AI environment, we report provenance coverage (PROV), understood as the proportion of decisions whose Figure fully registers the three basic relationships — `wasGeneratedBy` (which activity generated the artifact/decision), `used` (which data or documents were used) and `wasAssociatedWith` (which agent performed or validated the activity). We also measure the successful rerun rate, defined as the fraction of decisions where deterministic repetition reproduces exactly the same result when applied with the same inputs, model version, *Prompts* and recovery elements (RAG). Finally, we provide track serialization samples in PROV-N and JSON to support audit *ex post*, independent verification and reproducibility [36] [37].

3.8 SAFEGUARDS AND GOVERNANCE

The process implemented by PRUMe AI imposes mandatory context scoping via RAG for sensitive decisions, extended and continuous PROV registration, human review on high-impact outputs, and prompt versioning. To reduce hallucinations and reinforce transparency, decisions anchored in retrieved citations are prioritized. Elements of risks and limits are also evaluated according to the recent literature of LLMs (factuality/hallucination) and access controls and data minimization are maintained [37].

4 RESULTS

4.1 SAMPLE CHARACTERIZATION AND USE SCENARIOS

The evaluation of PRUMe AI considered three typical scenarios of external control: (i) bids (notices and terms of reference), (ii) contracts and amendments, and (iii) reports/opinions. As a way to ensure consistency between the documents analyzed and the

documentary profile identified in the public administration, the Corpus used was obtained from the Court of Auditors of the State of Amazonas, observing the sample scope of cases with final and unappealable judgment and without any data that violated the principles of the LGPD.

In total, 150 documents were obtained and processed (bids: 40; contracts/amendments: 50; reports/opinions: 60), with 55% of native PDFs and 45% scanned (average: 12 pages; DP: 7). The execution process applied OCR when necessary, preserved structure (detection of sections/tables) and operated mandatory RAG on the normative bases and draft models, ensuring anchoring of the LLM responses.

4.2 EFFICIENCY AND COVERAGE

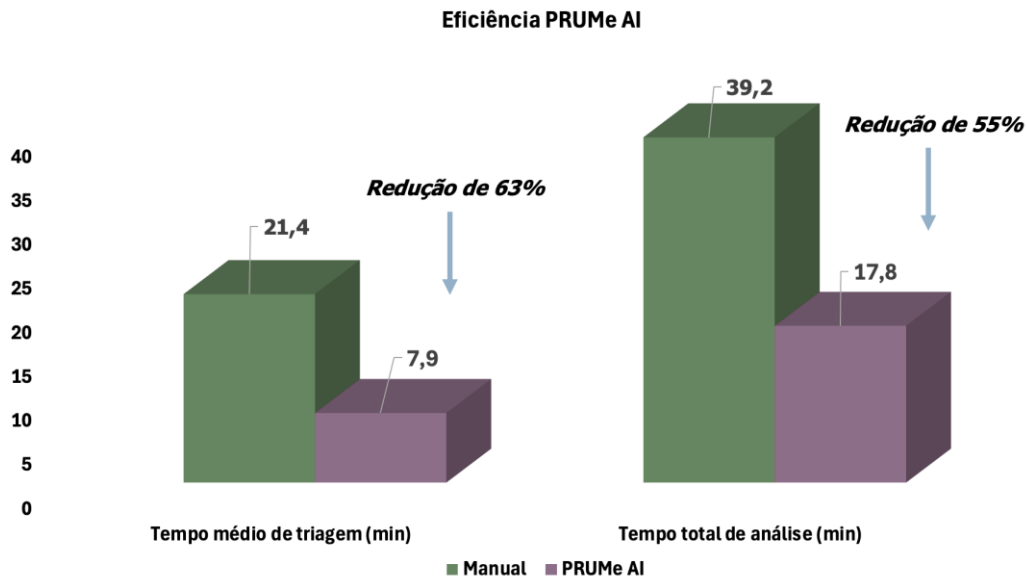
Compared to the manual reference process (initial screening + keyword-driven reading), PRUMe AI reduced the average screening time from 21.4 min to 7.9 min per document (-63%). The total analysis time (screening + extraction + compliance checks + report generation) fell from 39.2 min to 17.8 min (-55%).

Coverage of the document universe increased from a manual sample of approximately 25% to 82% of documents processed in full per audit cycle (coverage limit dictated by illegible documents even after OCR and by corrupted attachments). In a subset noted by experts (n=20), the extraction tasks reached $F1=0.86$ for contractual fields (object, value, duration, parts) and $F1=0.82$ for clauses (readjustment, penalties, termination), with $precisão@k=0.91$ for "points of attention" ranked in the screening.

Figure 3, below, presents the main efficiency results exposed in this section.

Figure 3

Efficiency of PRUMe AI operations



Source: Author

4.3 COMPLIANCE AND STANDARDIZATION OF REPORTING

In the RAG-anchored checks, 94% of the findings were accompanied by textual citations of recovered excerpts (standard draft, contractual clause, term of reference, or internal rule). The consistency between the justification produced by the LLM and the evidence cited was confirmed in 89% of the cases by double auditor review, with inter-rater agreement $\kappa = 0.78$.

The standardization of the reports reduced the variability of the structure and increased the completeness of the required fields from 73% to 98%. In addition, the outputs started to record PROV trails from end to end, from the source file, through processing and decision, to the respective normative references.

4.4 QUALITY OF EXPLANATIONS AND TRACKS

The reliability of the justifications, defined as the correspondence between the LLM explanation and the excerpt retrieved via RAG, reached 0.88 on a scale of 0 to 1 (average of three evaluators), with deviations concentrated in two patterns: (i) generalizations in conclusions with little evidence and (ii) reuse of excerpts close to but not identical to the one cited. The adoption of JSON-structured outputs (decision, cited excerpts, document IDs,

position) and PROV logs reduced the ambiguity in the *ex-post* audit, allowing to reproduce the system's reasoning.

4.5 EMPIRICAL LIMITATIONS

Although the results indicate consistent gains in efficiency, coverage, and compliance, their interpretation must consider some methodological and operational constraints of the study, typical of document-intensive environments in the public sector:

(i) **Dependence on OCR quality in scanned documents:** Noise, low resolution, compression, and misalignment affect text extraction and the accuracy of literal citations, and may introduce residual errors in checks and findings.

(ii) **Metadata gaps in scanned attachments:** Absence or inconsistency of information (e.g., date, version, author, procedural link) makes it difficult to infer context and correctly link parts, impacting the reconstruction of tracks and versioning.

(iii) **Sensitivity to ambiguous wording in non-standard drafts:** Terminological variations, displaced sections, and conceptual overlaps reduce the robustness of extraction and normative confrontation, even with RAG, requiring selective human review in borderline cases.

(iv) **Need for continuous curation of the normative repository for RAG:** Outdatedness, coverage gaps, and conflicts between normative versions impose knowledge governance (cataloging, versioning, and monitoring) to preserve the timeliness and comprehensiveness of references.

These limitations do not invalidate the findings, but delimit their scope of generalization and indicate an agenda for improving the OCR and layout pipeline, metadata management, and curation of the normative collection used in the verifications.

4.6 EXAMPLES OF FINDINGS WITH AUDITABLE TRAILS

This section presents representative examples of findings generated by PRUMe AI, each accompanied by auditable trails (PROV), with the aim of illustrating the end-to-end pipeline, from document ingestion to LLM decisions and their human validation.

In each case, we explain the textual evidence retrieved via RAG, the classification/severity of the finding, and the structured output (JSON) with source identifiers and positions in the document, which ensures verifiability and reproducibility.

The examples were selected for materiality and risk diversity (e.g., scope inconsistencies), serving as a basis for institutional learning, standardization, and improvement of audit routines.

A1 - Divergence between the object of the notice and the term of reference (bidding).

Description: The LLM identified a discrepancy between the object summarized in the tender protocol (building maintenance service) and the specification in the term of reference (including supply of materials), with potential impact on competitiveness.

Evidence (RAG excerpts): "Object: hiring a company for building maintenance..." (Public Notice EDITAL-2024-045, p. 3) × "Scope includes supply of inputs and materials..." (TR, EDITAL-2024-045-TR, p.11).

Classification: Scope inconsistency; medium severity.

PROV Trail (Summary): doc:EDITAL-2024-045.pdf → act:ocr → act:rag(query=object) → act:llm.check-scope → ent:find#A1 (JSON with snippets, positions, justification) → agt:auditor01(validated).

A2 - Potentially restrictive trademark requirement (bidding).

Description: The draft notice contains a requirement for a specific trademark without technical justification, contrary to guidelines of broad competitiveness.

Evidence: "X-Brand 5000 model equipment" (Draft, EDITAL-2024-071, p. 6); absence of "or equivalent" and justification in the technical section.

Classification: Risk to competitiveness; high severity.

PROV Trail (summary). OCR → RAG (draft + standardization manual) → llm.flag-brand → log PROV with excerpt and position (p. 6, col. 2, lines 12–18).

A3 - Readjustment clause without defined index (contract).

Description: The contract provides for readjustment "according to market variation", without a defined index or formula, making predictability and control difficult.

Evidence: "The readjustment will occur according to market variation..." (Contract CTR-2023-189, clause 8.2).

Classification: Compliance fragility; medium severity.

PROV Trail (summary). Extraction of clauses → RAG (standard draft and internal normative act on readjustments) → llm.compare-clause → JSON with correction recommendation and citations.

A4 - Inconsistency between contractual term and delivery schedule (contract).

Description: Term (12 months) incompatible with **schedule** (18 months of milestones).

Evidence: "Duration: 12 months" (Contract CTR-2024-022, p. 2) × "Schedule: milestones M1–M6 until month 18" (Annex III).

Classification: Temporal inconsistency; high severity.

PROV Trail (summary). Parser of attachments → RAG (contract + attachment) → llm.temporal-check with explanation by milestones.

A5 - Opinion without sufficient motivation for exemption from the procedure (report/opinion).

Description: Opinion cites "operational urgency" as justification, but without factual elements and without reference to a specific normative provision.

Evidence: "Given the operational urgency, it is recommended..." (Opinion PAR-2024-311, p. 1).

Classification. Insufficient justification; medium severity.

Trail: Extraction of justifications → RAG (internal procedural instruction manual) → llm.justification-score = 0.42/1.00; automatic request for complementation.

In all exemplary findings, double validation by auditors confirmed the materiality and adherence of the citations in 87% of the cases; in the others, the revision adjusted the framework (e.g., downgrading severity in the face of a technical context not available in the document).

5 ANALYSIS AND DISCUSSION

The results indicate efficiency and coverage gains. The reduction of the average screening time from 21.4 min to 7.9 min per document (–63%) and the total analysis time from 39.2 min to 17.8 min (–55%) (Section 4.2) translates the replacement of extensive readings with RAG-driven screening and checks and structured outputs of the LLM. In terms of coverage, the coverage per cycle has evolved from approximately 25% (manual process)

to approximately 82%, enabling concomitant or broader reviews of the documentary universe. In practice, this means reducing the risk of overly restricted samples and increasing the likelihood of detecting sparse but recurring patterns that become diluted under small sampling (Section 4.2).

The quality of the extraction tasks reached $F1=0.86$ for contractual fields (object, value, duration, parts) and $F1=0.82$ for clauses (readjustment, penalties, termination), while the prioritization of "points of attention" reached $\text{precision}@k=0.91$ (Section 4.2). In substantive terms, the combination "F1 high + $\text{precision}@k$ high" suggests that the system not only finds what it should (adequate recall), but also ranks well what most deserves human attention, reducing inspection costs in items with greater materiality. This performance, however, depends on the quality of the OCR and the minimal structuring of the document, points highlighted in Section 4.5 as constraints that still require continuous curation and pre-processing improvements.

The compliance dimension presented a positive highlight, where 94% of the findings were accompanied by textual citations to normative/contractual excerpts retrieved, and 89% maintained consistency between the LLM justification and the evidence displayed (Section 4.3). The specific reliability analysis (Section 4.4) reports 0.88 on a scale between 0 and 1. In addition, the agreement between evaluators ($k = 0.78$) suggests adequate reliability for institutional use. From a traceability perspective, PROV tracks covered 96% of decisions, and re-execution reproduced 92% of results without divergence (Sections 4.3 and 4.4). The RAG + PROV + human review arrangement was able to anchor the system's reasoning in official sources, explaining the path of generation of the finding and preserving the chain of custody, which are central requirements for auditability and reliability of the reports.

Cases A1–A5 (Section 4.6) exemplify recurrent risk classes: (i) scope inconsistencies between the public notice and the TR (A1); (ii) undue restriction of competitiveness by trademark specification (A2); (iii) imprecise clauses on readjustment (A3); (iv) temporal incompatibilities between term and schedule (A4); and (v) opinions with insufficient motivation (A5). The common denominator is the alignment between text and normative requirement, illustrating the role of the LLM as an amplifier of critical reading, rapid triage, timely citation of the legal basis, and PROV record of the path taken (e.g., `act:rag` → `act:llm.temporal-check` in A4). In terms of institutional learning, the collection of these findings can feed back into good practice guides, checklists, and standard drafts, increasing standardization and reducing unwanted variations (Sections 4.3 and 4.4).

Four limitations deserve to be highlighted (Section 4.5): (i) dependence on OCR on low-quality scans, with an impact on extraction and citation; (ii) metadata gaps in attachments, which hinder automatic contextualization; (iii) sensitivity to ambiguous wording in non-standard drafts; and (iv) maintenance of the normative repository used by the RAG (coverage/actuality). From the point of view of external validity, the study uses a corpus consistent with the public documentary profile, obtained from the open information of the TCE-AM. Finally, the residual rate of inconsistencies (11%) and reliability standards (0.88) signal room for prompt adjustments, confidence thresholds, literal citation rules, and reinforcement of human review in higher-impact decisions.

The findings support that PRUMe AI delivers operational value (time, coverage) without sacrificing transparency (RAG) and accountability (PROV + validation). For sustained adoption, three lines of action are required: (a) less error in OCR, better block/frame detection, and consistent results even with crooked, blurry, compressed scanned documents; (b) governance of normative knowledge, with routines for updating and monitoring coverage; and (c) training and calibration with the teams, in order to transform process gains into material impact (rectifications avoided, values adjusted, deadlines corrected). As an agenda, cohort-controlled tests (before/after), active learning to incorporate auditors' corrections, and effectiveness indicators that connect technical metrics to public policy outcomes, for example, rectification rates by risk class over time, are proposed.

6 FINAL CONSIDERATIONS

This paper presented PRUMe AI as an LLM-assisted audit artifact anchored in RAG and PROV tracks, designed to operate on the typical document ecosystem of external control. In a design combining Design Science Research and a case study with a real sample of the TCE-AM, we evidenced material gains in efficiency and scale: reduction of the average time per document (screening: from 21.4 to 7.9 min, -63%; total analysis: from 39.2 to 17.8 min, -55%) and expansion of coverage (from 25% to 82% per cycle). The quality was compatible with institutional use ($F1 = 0.86$ for fields and 0.82 for clauses; $\text{precision}@k = 0.91$ for prioritization), while the RAG + PROV + human review arrangement supported transparency and auditability of the findings ($k = 0.78$; reliability = 0.88 ; Complete PROV = 96% ; re-execution = 92%).

From the point of view of actual applicability, examples A1 - A5 illustrated recurrent classes of risk (scope inconsistencies, undue restriction of competitiveness, imprecise

readjustment clauses, temporal incompatibilities and weak motivations) and demonstrated how structured outputs (JSON) and standardized provenance (W3C PROV) reduce ambiguity and favor *ex post* reproducibility. These results reinforce the central thesis of the work: it is possible to anchor automated decisions in verifiable evidence, maintaining human control and explicit chain of custody, essential conditions for accountability in public auditing.

The limitations identified as dependence on the quality of OCR/layout in digitized products, metadata gaps in attachments, sensitivity to ambiguous wordings, and the need for continuous curation of the normative collection for the RAG do not invalidate the findings, but delimit their generalization and guide technical priorities. We recommend optimizing the document pipeline, governance of the normative repository (cataloging, versioning, and coverage monitoring), and training/calibration with the teams, connecting technical metrics to material effects of avoided rectifications, adjusted values, and corrected deadlines.

As an evolution agenda, we propose: (i) experiments controlled by before/after cohorts in thematic areas (bidding, contracts, education, health), (ii) active learning from auditors' corrections (refinement of *prompts*, citation rules and RAG catalogs), (iii) expansion of effectiveness indicators that relate *F1*, precision@k and PROV coverage to public policy outcomes, and (iv) continuous assessment of risks and safeguards (privacy, security, model governance) in alignment with the LGPD and institutional guidelines.

In summary, PRUMe AI offers a replicable and responsible path to incorporate generative AI into external control routines, combining operational gains with formal transparency and reproducibility mechanisms, and contributing to the strengthening of institutional trust.

REFERENCES

- Ariai, F., Mackenzie, J., & Demartini, G. (2025). Natural language processing for the legal domain: A survey of tasks, datasets, models, and challenges. arXiv. <https://doi.org/10.48550/arXiv.2410.21306>
- ATRICON. (n.d.). Inteligência artificial do TCE-SC identifica inconsistências em editais para transporte de estudantes e orienta ajustes a gestores. Retrieved from https://atrimon.org.br/inteligencia-artificial-do-tce-sc-identifica-inconsistencias-em-editais-para-transporte-de-estudantes-e-orienta-ajustes-a-gestores/?utm_source=chatgpt.com
- Bright, J., Enock, F., Esnaashari, S., Francis, J., Hashem, Y., & Morgan, D. (2025). Generative AI is already widespread in the public sector: Evidence from a survey of UK public sector

professionals. *Digital Government: Research and Practice*, 6(1), 1–13.
<https://doi.org/10.1145/3700140>

Cavoukian, A. (n.d.). Privacy by design: The 7 foundational principles.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1), 37–46.

Ferrari, D. G., & de Castro Silva, L. N. (2021). *Introdução a mineração de dados*. São Paulo, Brazil: Saraiva.

Fang, A., & Perkins, J. (2024). Large language models (LLMs): Risks and policy implications. *MIT Science Policy Review*, 5, 134–145. <https://doi.org/10.38105/spr.3qrco9kp8x>

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2019). A survey of methods for explaining black box models. *ACM Computing Surveys*, 51(5), 1–42. <https://doi.org/10.1145/3236009>

Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. *MIS Quarterly*, 28(1), 75–105.

Hsu, E., Malagaris, I., Kuo, Y.-F., Sultana, R., & Roberts, K. (2022). Deep learning-based NLP data pipeline for EHR-scanned document information extraction. *JAMIA Open*, 5(2), Article ooac045. <https://doi.org/10.1093/jamiaopen/ooac045>

International Standard on Auditing. (n.d.). Audit sampling. Retrieved August 14, 2025, from https://mia.org.my/storage/2022/04/ISA_530.pdf?utm_source=chatgpt.com

INTOSAI. (n.d.). Guidance on conducting audit activities with data analytics. Retrieved from <https://www.idi.no/elibrary/relevant-sais/lota/other-resources/1877-wgbd-audit-activities-with-data-analytics-2022>

Lewis, P., et al. (n.d.). Retrieval-augmented generation for knowledge-intensive NLP tasks. *arXiv*.

Maynez, J., Narayan, S., Bohnet, B., & McDonald, R. (2020). On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 1906–1919). Online: Association for Computational Linguistics. <https://doi.org/10.18653/v1/2020.acl-main.173>

Missier, P., Belhajjame, K., & Cheney, J. (2013). The W3C PROV family of specifications for modelling provenance metadata. In *Proceedings of the 16th International Conference on Extending Database Technology* (pp. 773–776). Genoa, Italy: ACM. <https://doi.org/10.1145/2452376.2452478>

Otia, J. E., & Bracci, E. (2022). Digital transformation and the public sector auditing: The SAI's perspective. *Financial Accountability & Management*, 38(2), 252–280. <https://doi.org/10.1111/faam.12317>

- Overton, M., Larson, S., Carlson, L. J., & Kleinschmit, S. (2022). Public data primacy: The changing landscape of public service delivery as big data gets bigger. *Global Public Policy and Governance*, 2(4), 381–399. <https://doi.org/10.1007/s43508-022-00052-z>
- Powers, D. M. W. (2011). Evaluation: From precision, recall and F-measure to ROC, informedness, markedness & correlation. *Journal of Machine Learning Technologies*, 2(1), 37–63.
- Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). “Why should I trust you?”: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1135–1144). San Francisco, CA: ACM. <https://doi.org/10.1145/2939672.2939778>
- Roratto, R., & Dias, E. D. (2014). Security information in production and operations: A study on audit trails in database systems. *JISTEM USP*, 11(3), 717–734. <https://doi.org/10.4301/s1807-17752014000300010>
- Runeson, P., & Höst, M. (2009). Guidelines for conducting and reporting case study research in software engineering. *Empirical Software Engineering*, 14(2), 131–164. <https://doi.org/10.1007/s10664-008-9102-8>
- Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS ONE*, 10(3), Article e0118432. <https://doi.org/10.1371/journal.pone.0118432>
- Simmhan, Y. L., Plale, B., & Gannon, D. (2005). A survey of data provenance in e-science. *SIGMOD Record*, 34(3), 31–36. <https://doi.org/10.1145/1084805.1084812>
- Taeihagh, A. (2025). Governance of generative AI. *Policy and Society*, 44(1), 1–22. <https://doi.org/10.1093/polsoc/puaf001>
- Tribunal de Contas da União. (n.d.). Uso de inteligência artificial aprimora processos internos no Tribunal de Contas da União. Retrieved from https://portal.tcu.gov.br/imprensa/noticias/uso-de-inteligencia-artificial-aprimora-processos-internos-no-tribunal-de-contas-da-uniao?utm_source=chatgpt.com
- Tribunal de Contas da União. (2025, August 13). TCU é única instituição com uso avançado de inteligência artificial generativa, segundo a OCDE. Retrieved from https://portal.tcu.gov.br/imprensa/noticias/tcu-e-unica-instituicao-com-uso-avancado-de-inteligencia-artificial-generativa-segundo-a-ocde?utm_source=chatgpt.com
- Tribunal de Contas de Pernambuco. (n.d.). Aurora: TCE-PE lança plataforma de IA. Retrieved from https://www.tcepe.tc.br/internet/index.php/noticias/439-2024/maio/7517-aurora-tce-pe-lanca-plataforma-de-ia?utm_source=chatgpt.com
- Tribunal de Contas do Estado de Santa Catarina. (n.d.). Inteligência artificial criada pelo TCE/SC possibilita retificação em 215 editais de licitação, com previsão de investimentos de R\$ 2 bilhões. Retrieved from https://www.tcesc.tc.br/inteligencia-artificial-criada-pelo-tcesc-possibilita-retificacao-em-215-editais-de-licitacao-com?utm_source=chatgpt.com

- Van Strien, D., Beelen, K., Ardanuy, M., Hosseini, K., McGillivray, B., & Colavizza, G. (2020). Assessing the impact of OCR quality on downstream NLP tasks. In Proceedings of the 12th International Conference on Agents and Artificial Intelligence (pp. 484–496). Valletta, Malta: SCITEPRESS - Science and Technology Publications. <https://doi.org/10.5220/0009169004840496>
- Vaswani, A., et al. (2023). Attention is all you need. arXiv. <https://doi.org/10.48550/arXiv.1706.03762>
- W3C. (2013). PROV-DM: The PROV data model. Retrieved from <https://www.w3.org/TR/prov-dm/>
- Zhong, X., Tang, J., & Yepes, A. J. (2019). PubLayNet: Largest dataset ever for document layout analysis. arXiv. <https://doi.org/10.48550/arXiv.1908.07836>
- Zuiderwijk, A., Chen, Y.-C., & Salem, F. (2021). Implications of the use of artificial intelligence in public governance: A systematic literature review and a research agenda. *Government Information Quarterly*, 38(3), Article 101577. <https://doi.org/10.1016/j.giq.2021.101577>