Revista ARACÊ
ISSN: 2358-2472

CONTROLE EXTERNO BASEADO EM LLM: AVALIAÇÃO EMPÍRICA DO PRUME AI

LLM-BASED EXTERNAL CONTROL: EMPIRICAL EVALUATION OF PRUME AI

CONTROL EXTERNO BASADO EN LLM: EVALUACIÓN EMPÍRICA DE PRUME AI

https://doi.org/10.56238/arev7n9-271

Data de submissão: 25/08/2025 **Data de publicação:** 25/09/2025

Alessandro de Souza Bezerra
Doutorado
Instituição: Universidade do Estado do Amazonas
E-mail: abezerra@uea.edu.br
Orcid: https://orcid.org/0000-0002-6410-7099

Luciane Cavalcante Lopes
Mestre
Instituição: Tribunal de Contas do Estado do Amazonas
E-mail: luciane.lopes@tce.am.gov.br
Orcid: https://orcid.org/0009-0004-9449-5661

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 (campos contratuais) e F1=0,82 (cláusulas), com 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). Concluímos 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.

Revista ARACÊ
ISSN: 2358-2472

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.

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 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.

Revista ARACÊ
ISSN: 2358-2472

1 INTRODUÇÃO

As atividades de auditoria e controle público enfrentam um cenário de transformação contínua impulsionado pela digitalização de processos e pela massificação de dados administrativos. Contratos, notas fiscais, relatórios de gestão e comunicações oficiais são gerados em volume cada vez maiores e formatos amplamente diversos, os quais exigem métodos capazes de lidar com heterogeneidade e escala. Embora o avanço do processo de digitalização do setor público tenha ampliado a disponibilidade de evidências, o aumento da complexidade e diversidade informacional tende a superar a capacidade de análise manual, resultando em gargalos de avaliação, assimetrias de cobertura e variações na qualidade dos achados [1].

O desequilíbrio entre oferta de dados e capacidade analítica cria lacunas de tempo e de escala. No tempo, a validação tardia reduz a tempestividade do controle e limita a possibilidade de correção antes da consolidação de danos. Na escala, a definição de amostras fortemente restritas impõe riscos de não detecção de padrões de irregularidade que se manifestam de modo esparso, porém recorrente. Ademais, a fragmentação das trilhas de auditoria e a ausência de padronização na documentação dos achados dificultam a reprodutibilidade das análises e a prestação de contas [2].

Os Tribunais de Contas no Brasil vêm adotando soluções inovadoras baseadas em Inteligência Artificial (IA) para dar escala e tempestividade ao controle externo, com evidências recentes de difusão e amadurecimento institucional. Levantamento do IRB/Atricon indica que, em 2024, o uso de IA no Controle Externo cresceu de 18 para 28 tribunais, movimentando também capacitação e governança interna; no plano federal, a OCDE classificou o TCU como caso de uso avançado de IA generativa no setor público, destacando a oferta institucional da tecnologia aos servidores [3].

Em uma perspectiva prática, despontam iniciativas voltadas diretamente à fiscalização de editais e processos. O TCE-SC desenvolveu o sistema VigIA, que analisa licitações antes da publicação e sinaliza inconsistências para ação concomitante: entre 18/4 e 8/10 (2024), 7.711 editais foram processados, gerando 63.445 respostas automáticas e resultando em 215 correções (revogações/retificações etc.), com materialidade na ordem de R$ 2 bilhões [4]. Também em Santa Catarina, notas públicas e matérias setoriais reforçam o uso do VigIA para transporte escolar e outras contratações, ampliando a seleção de riscos a serem observados e aprofundados por auditores [5].

Outros tribunais vêm estruturando plataformas generativas para apoiar análise documental e atendimento a jurisdicionados. O TCE-PE lançou a plataforma Aurora (2024) e, em 2025, divulgou evolução do AuroraChat com novas funções para sumarização e extração de informações processuais; o TCE-PR opera o AVIA (atendimento virtual por IA) e o ChatTCEPR (chat institucional com LLM), ambos com foco em produtividade interna e serviço ao fiscalizado; o TCU disponibiliza o ChatTCU

Revista ARACÊ
ISSN: 2358-2472

e publicou guia de uso responsável de IA generativa para orientar salvaguardas, transparência e conformidade [6] [7].

Diante do contexto apresentado, propomos o PRUMe AI, uma plataforma web de controle e auditoria assistida que combina técnicas de Processamento de Linguagem Natural e mecanismos de explicabilidade para apoiar equipes de controle no exame sistemático de grandes volumes documentais. A plataforma produz trilhas auditáveis e relatórios estruturados, priorizando transparência, reprodutibilidade e aderência às regulações aplicáveis. Diferentemente de abordagens puramente manuais ou caixas-pretas, o PRUMe AI prioriza a interpretabilidade dos achados e a rastreabilidade das evidências.

O objetivo geral deste estudo é apresentar a arquitetura técnica e funcional do PRUMe AI, bem como os resultados práticos que evidenciam sua capacidade de (i) aumentar a eficiência da atividade de auditoria, por meio da redução do tempo de triagem e análise; (ii) ampliar a cobertura, possibilitando o exame de universos documentais mais extensos; e (iii) fortalecer a conformidade, por meio da checagem sistemática de requisitos normativos e da padronização de relatórios.

2 FUNDAMENTAÇÃO TEÓRICA

2.1 CONTROLE EXTERNO NA ERA DIGITAL

Os Tribunais de Contas operam em um ecossistema informacional intensivo em dados, no qual a efetividade do controle depende da capacidade institucional de transformar grandes massas de registros administrativos em evidência auditável com qualidade e tempestividade. A literatura recente em administração pública e governo digital aponta que a adoção de IA e analytics reconfigura rotinas de trabalho, cria novas capacidades de monitoramento e impõe salvaguardas de transparência — movimento que exige governança de dados, competências e padrões para decisões baseadas em evidências [8].

Nesse contexto, cresce o emprego de data analytics no setor público, com relatos empíricos de uso em órgãos de controle e administração direta, indicando que a expansão do “big data” altera práticas de fiscalização e desenho de serviços, mas também expõe desafios de qualidade de dados, interoperabilidade e mensuração de valor público [9].

Para o controle externo, o resultado prático é a iminente necessidade de incorporar ciência de dados e analytics aos ciclos de auditoria, com capacitação continuada e padrões de trabalho que viabilizem a análise em larga escala e a revisão baseada em risco.

Revista ARACÊ
ISSN: 2358-2472

2.2 AUDITORIA BASEADA EM DADOS E AUDITORIA CONTÍNUA

A auditoria baseada em dados consolida o uso de técnicas de analytics ao longo do ciclo de auditoria (planejamento, execução e apresentação de resultados), permitindo tanto testes sobre populações completas quanto amostragem estatística apropriada. Em termos normativos, a amostragem permanece relevante para obtenção de evidência suficiente e apropriada; contudo, a disponibilidade de dados e automação amplia a viabilidade de análises exaustivas em áreas críticas [10].

A literatura sobre auditoria contínua descreve a migração de procedimentos periódicos para rotinas automatizadas e próximas do evento, com alertas e métricas operacionais integradas aos processos—um paradigma que reduz gargalos temporais e favorece a detecção precoce. Estudos de caso e sínteses acadêmicas detalham princípios, benefícios e desafios de implementação, inclusive a necessidade de reengenharia de processos e de automação como fundamento [11].

Para o setor público, diretrizes específicas da comunidade INTOSAI orientam a condução de atividades de auditoria com data analytics, cobrindo conceito, processo, competências e mecanismos de trabalho, e reforçando a integração com a avaliação de risco [12].

2.3 PROCESSAMENTO DE LINGUAGEM NATURAL (PLN) EM DOCUMENTOS PÚBLICOS

Em cenários relacionados a licitações, contratos, relatórios e outros documentos administrativos e oficiais, o PLN viabiliza tarefas de classificação documental, extração de entidades e cláusulas, correspondência semântica e sumarização, compondo um pipeline típico de ingestão, OCR, extração/normalização e indexação/pesquisa. Pesquisas recentes no escopo jurídico descrevem os desafios do texto legal (comprimento, linguagem técnica, escassez de dados abertos) e mapeiam tarefas e modelos aplicáveis [13].

A qualidade do OCR e os ruídos característicos de digitalizações afetam diretamente o desempenho de tarefas *downstream*, como NER e classificação, exigindo cuidados de pré-processamento e curadoria para preservar a rastreabilidade da evidência [14]. De forma contínua, pipelines documentais em outros escopos regulados reforçam a sequência OCR e mineração de textos como base técnica para minerar registros digitalizados em escala, experiência totalmente aplicável ao setor público [15].

2.4 IA GENERATIVA E MODELOS DE LINGUAGEM NO SETOR PÚBLICO

A IA generativa, apoiada por Modelos de Linguagem de grande porte (LLMs), tem sido incorporada a rotinas de governo para redação assistida, análise de documentos, atendimento ao

Revista ARACÊ
ISSN: 2358-2472

cidadão e suporte a tomada de decisão. Evidências empíricas recentes mostram adoção ampla no setor público: um inquérito com profissionais do serviço público no Reino Unido aponta uso disseminado de ferramentas generativas e percepção de ganhos de produtividade e redução de carga burocrática, mas também lacunas de diretrizes institucionais para seu uso. Esses achados convergem com a literatura especializada que enxerga a tecnologia como infraestrutura de suporte a serviços e back-office, demandando arranjos de governança e competências específicas [16].

No contexto organizacional, a adoção de IA no setor público envolve **elementos de** eficiência, equidade e transparência, embora suscite desafios **de implementação** relacionados a centralização e experimentação de soluções. Na perspectiva técnico-metodológica, LLMs ampliam o leque de aplicações em "Governo Inteligente", mas impõem desafios de veracidade, vieses e segurança. A literatura de LLMs aplicados ao setor público enfatiza o desenho de pipelines com *Retrieval-Augmented Generation (RAG)* para fundamentar respostas em documentos oficiais, reduzir alucinações e viabilizar auditabilidade das saídas, ao tempo que discute o equilíbrio entre desempenho e interpretabilidade. Análises de risco e de políticas públicas sobre LLMs destacam implicações regulatórias - direitos autorais, privacidade e transparência algorítmica - e recomendam métricas e trilhas de verificação adequadas a escopos regulados [17].

A agenda de governança para IA generativa na administração pública tem avançado na tentativa de acompanhar a velocidade de adoção da tecnologia. Estudos recentes sistematizam os riscos de alucinação, *jailbreaking*, vazamento de dados e manipulação, adotando mecanismos de mitigação como avaliação de impacto, *guardrails* e auditorias. Tais iniciativas apontam para necessidade de capacidade regulatória e de responsabilização no uso de LLMs [18]. Em tribunais de contas, controladorias e reguladores, isso se traduz na combinação de políticas internas, compliance técnico (registros de proveniência, logs, restrições de acesso) e processos de revisão humana para preservar a legitimidade das decisões.

Finalmente, a literatura sinaliza que a captura de valor público com IA generativa depende menos de "provas de conceito" isoladas e mais de capacidade institucional dirigida a dados governamentais de qualidade, processos redesenhados para auditoria e mensuração de resultados, e mecanismos de aprendizado organizacional. Estudos sobre adoção estratégica de IA em governos reforçam que o caminho crítico passa por governança de dados, gestão de competências e integração com métricas de desempenho do serviço público, condições que, quando atendidas, permitem que LLMs e RAG operem como infraestrutura de eficiência, cobertura e padronização nas rotinas de gestão e controle [19].

Revista ARACÊ
ISSN: 2358-2472

2.5 EXPLICABILIDADE, TRILHAS AUDITÁVEIS E CONFORMIDADE

A adoção responsável de IA na auditoria requer **explicabilidade** (XAI) para sustentar confiança, capacidade de revisão e prestação de contas. Pesquisas anteriores sistematizam métodos de explicação e discutem o equilíbrio entre interpretabilidade e desempenho. Adicionalmente, técnicas como **LIME** ilustram abordagens que viabilizem ajustes ágeis e facilitados para justificar previsões em tarefas de texto [20] [21].

No plano das **trilhas auditáveis** e da reprodutibilidade, padrões abertos de **proveniência**, como o modelo **W3C PROV-DM**, permitem registrar entidades, atividades e agentes envolvidos na geração de artefatos (dados, modelos, relatórios), fortalecendo a cadeia de manutenção de evidência. Em sistemas informacionais, grupos de controles de **auditoria e responsabilização** orientam geração, proteção e revisão de registros de log como salvaguarda técnica [22].

Quanto à **proteção de dados**, a **LGPD (Lei nº 13.709/2018)** estabelece princípios e bases legais para o tratamento de dados por órgãos públicos, demandando desenho de soluções com **desenhos de proteção de dados e privacidade** através da minimização, retenção e segurança, partindo da fase inicial do ciclo de vida da informação. Guias clássicos de **modelos de privacidade** oferecem princípios operacionais compatíveis com tais exigências [23].

3 MÉTODO

3.1 ARQUITETURA DO PRUME AI

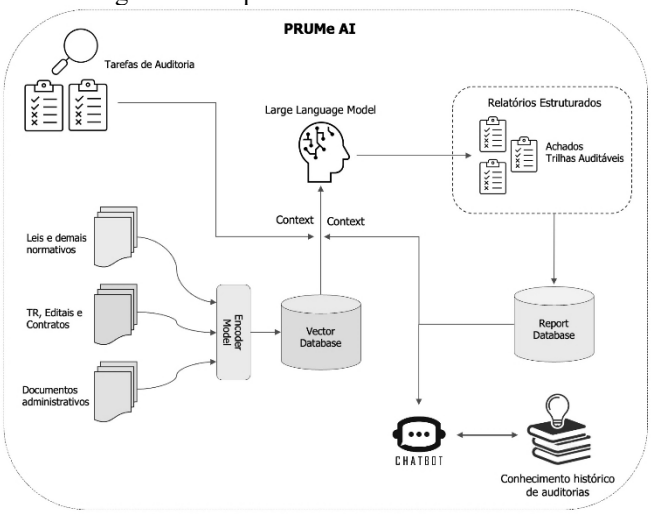
A Figura 1 apresenta a arquitetura geral do PRUMe AI, a qual pode ser apresentada por meio de 3 blocos funcionais principais: (i) Entrada e tratamento de dados de auditoria; (ii) Análise de dados de auditoria baseada em LLM; e (iii) Apresentação de resultados e consulta de conhecimento histórico de auditorias.

Inicialmente, documentos de auditoria – termos de referência, legislações, normas gerais, editais, contratos e documentos administrativos - são carregados na plataforma PRUMe AI para que sirvam de contexto de análises pelo LLM. Adicionalmente, prompts orientados ao processo de análise e identificação de achados de auditoria são submetidos como complemento contextual.

O processo de análise de dados é realizado na camada de análise de dados de auditoria por LLM, processando todos os documentos enviados sob as diretrizes indicadas nos prompts de contexto. Por fim, na camada funcional de apresentação de resultados e consulta de conhecimento histórico de auditorias são apresentados os resultados gerados pelo PRUMe AI, além da disponibilização de um mecanismo de Chat capaz de realizar consultas e gerar respostas baseadas no conhecimento histórico de auditorias já processadas pelo PRUMe AI.

Revista ARACÊ
ISSN: 2358-2472

Figura 1 – Arquitetura Geral do PRUMe AI



Fonte: Autor

3.2 ABORDAGEM DE PESQUISA

De maneira generalista adotamos um processo de modelagem centrado na construção e avaliação de artefato, combinando Design Science Research (DSR) para a engenharia da plataforma e um estudo de caso para observação e avaliação em uso real. A DSR orientou o ciclo iterativo problema por meio das etapas de: requisitos, projeto/artefato, demonstração, avaliação e refinamento, assegurando alinhamento entre necessidades de auditoria e decisões técnicas e de arquitetura do PRUMe AI [24].

A avaliação em contexto real do PRUME AI ocorreu por meio de estudo de caso específico, com múltiplas unidades de análise (editais, contratos e relatórios), protocolo explícito de coleta e triangulação (logs do sistema, amostras documentais e julgamento de auditores), seguindo boas práticas de delineamento e reporte em engenharia de software [25].

3.3 CORPUS E PREPARAÇÃO

O corpus (dados/documentos utilizados no processo de análise e valiação do PRUMe AI) inclui editais, termos de referência, contratos/aditivos e relatórios em PDF nativo e digitalizado. Documentos escaneados passam por OCR (deskew, limpeza, reconhecimento) e todos os itens são submetidos à análise de layout (página/blocos/tabelas) para preservar suas estruturas semânticas. Como base técnica e de boas práticas, foi utilizada a arquitetura do Tesseract (ICDAR'07), o conjunto de dados PubLayNet (ICDAR'19) e o modelo LayoutLM (KDD'20) [27].

Revista ARACÊ
ISSN: 2358-2472

3.4 ARQUITETURA DE AUDITORIA ASSISTIDA POR LLM

Toda a lógica de triagem, extração, classificação, verificação normativa, geração de achados e elaboração de trilhas auditáveis é realizada por meio do LLM GPT-4, acessado via API da OpenAI. A escolha toma como base em três pilares técnicos determinantes: (a) a capacidade dos Transformers de modelar dependências de longo alcance; (b) o desempenho de modelos em larga escala para tarefas few-shot; e (c) o alinhamento por feedback humano característico de famílias Instruct/GPT, que favorece saídas úteis e controláveis [28]. Para **mitigar alucinações** e garantir fundamentação, o PRUMe AI utiliza o paradigma de **Retrieval-Augmented Generation (RAG)**: antes de cada decisão/achado, o sistema consulta bases internas de leis, editais e normas e **fornece ao LLM trechos recuperados** como contexto obrigatório de resposta. O prompt passado ao LLM exige **citações dos trechos recuperados** sempre que uma conclusão normativa é apresentada e uma resposta é gerada [29].

3.5 SAÍDAS ESTRUTURADAS E TRILHAS AUDITÁVEIS

As respostas do LLM são emitidas em JSON com campos de decisão, trechos citados, referências normativas e IDs/posições no documento. Cada passo (OCR, recuperação, chamada ao LLM, validação humana) é registrado segundo o W3C PROV: Entities (documento, trecho, achado), Activities (ocr, retrieve, llm.check, report) e Agents (serviço LLM, auditor). O uso de PROV-DM permite reconstruir a linhagem de cada achado (who/what/when/how) e dá suporte à verificação independente. A seção 3.7 deste trabalho fornece mais detalhes sobre o W3C PROV. Adicionalmente, a disponibilização das resposta em formato JSON nos possibilita a integração posterior dos resultados a plataformas e mecanismos variados que demandem simplesmente a leitura padronizada dos conteúdos de interesse [30].

3.6 TAREFAS REALIZADAS PELO LLM

Podemos organizar o uso do LLM GPT-4 pelo PRUMe AI em um conjunto de tarefas encadeadas, cada uma delas com entradas, procedimentos e saídas claramente especificadas. Todas as decisões são fundamentadas em recuperação prévia de evidências (RAG) e registradas em trilhas PROV, com revisão humana nos pontos de maior materialidade.

I. Triagem

O LLM classifica automaticamente o tipo de documento a ser analisado (edital, termo de referência, contrato, aditivo, parecer) e estima a prioridade de análise com base em rubricas explícitas de risco (restrição de competição, desalinhamento objeto–escopo, lacunas de cláusulas essenciais

Revista ARACÊ
ISSN: 2358-2472

etc.). Como entradas de dados são passados elementos de texto integral ou seções detectadas e metadados básicos. As saídas são retornadas na forma de rótulos de tipo, score de risco e lista de pontos de atenção com referência às passagens relevantes. Essa etapa reduz o gargalo inicial e direciona esforço humano para os itens de maior potencial de materialidade.

II. Extração

Para documentos classificados, o LLM executa extração orientada a esquema (formato JSON) contendo partes, objeto, valores, vigência, itens de fornecimento/serviço e cláusulas-alvo (reajuste, penalidades, rescisão, garantias, fiscalização). As entradas se concentram nas passagens textuais segmentadas e, quando aplicável, regiões de tabelas passíveis de interpretação pelo mecanismo de OCR. As saídas são da forma de dicionários normalizados com IDs e posições no documento, onde cada campo traz o trecho de origem que sustenta a extração. Essa estruturação permite análises comparáveis entre processos e facilita a verificação posterior.

III. Conformidade

A verificação de conformidade é realizada pelo próprio LLM, mas com RAG obrigatório. Antes de responder, o sistema recupera trechos de normas, minutas e atos internos pertinentes e os adiciona ao contexto do prompt. O modelo então confronta o conteúdo extraído com os requisitos aplicáveis, como por exemplo: a existência de índice de reajuste, cláusulas de sanção, prazos mínimos. Nas saídas para cada requisito é definido a decisão (atende/não atende/indeterminado), justificativa textual e citação do trecho normativo utilizado. Ao exigir citação de referências, a etapa torna a decisão verificável e auditável.

IV. Relato

Quando há não conformidades ou fragilidades, o LLM gera achados estruturados contendo o tipo (p. ex., restrição à competitividade, inconsistência temporal), descrição sintética, grau de gravidade, base legal (com referência às normas recuperadas via RAG), e sugestão de correção quando aplicável. Além do texto, o sistema produz explicações locais relatando por que aquele trecho sustenta a conclusão, e liga cada justificativa às evidências citadas no passo anterior. O objetivo é padronizar a redação, preservar rastreabilidade e facilitar a revisão por pares.

V. Consolidação

Finalmente, o PRUMe AI consolida os resultados em relatórios padronizados, agregando extratos dos documentos, tabelas-resumo e a lista de achados com suas respectivas citações. Em paralelo, registra-se a proveniência completa no padrão W3C PROV: documentos e trechos como Entities, etapas de processamento (OCR, recuperação, checagens, geração) como Activities, e

Revista ARACÊ
ISSN: 2358-2472

serviços/sujeitos envolvidos (LLM, auditor revisor) como Agents. Essa trilha permite sua reexecução, verificação independente e controle de qualidade sobre o ciclo decisório.

A Figura 2, a seguir, apresenta uma visão parcial do relatório consolidado com os resultados de uma análise realizada pelo PRUMe AI, sob a perspectiva das tarefas realizadas pelo LLM e descritas nessa seção.

Figura 2 – Visão parcial de relatório consolidado do PRUMe AI

PRUMe AI — Relatório de Saída Consolidado (Execução 19/08/25)

ACHADOS ESTRUTURADOS

A seguir, são apresentados os achados gerados pelo PRUMe AI, com descrição sintética e representação JSON estruturada para auditoria e reexecução.

A1 — Inconsistência de escopo (edital vs. TR)

Discrepância entre objeto do edital e escopo do termo de referência, com potencial impacto na competitividade.


```
{
  "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"
}
```


A2 — Exigência potencialmente restritiva de marca

Minuta menciona modelo/marca específica sem 'ou equivalente'.


```
{
  "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'"
}
```


Fonte: Autor

Revista ARACÊ
ISSN: 2358-2472

3.7 MÉTRICAS E AVALIAÇÃO DO PRUME AI

A verificação da qualidade de extração e identificação de documentos foi realizada por meio da observação das métricas de precisão, revocação e F1-score em tarefas de classificação/extração (campos e cláusulas).

Em problemas de classificação binária, predições podem ter quatro possíveis classes [31], são elas:

Verdadeiro positivo (VP): quando o método diz que a classe é positiva e, ao verificar a resposta, vê-se que a classe era realmente positiva;

Verdadeiro negativo (VN): quando o método diz que a classe é negativa e, ao verificar a resposta, vê-se que a classe era realmente negativa;

Falso positivo (FP): quando o método diz que a classe é positiva, mas ao verificar a resposta, vê-se que a classe era negativa;

Falso negativo (FN): quando o método diz que a classe é negativa, mas ao verificar a resposta, vê-se que a classe era positiva.

Suas equações são apresentadas a seguir:

- **Precisão** = $\frac{VP}{VP+FP}$, que avalia a quantidade de verdadeiros positivos sobre a soma de todos os valores positivos;
- **Revocação** = $\frac{VP}{VP+FN}$, que avalia a capacidade do método de detectar com sucesso resultados classificados como positivos;
- **F1** = 2 * $\frac{\text{Precisão}*\text{Revocação}}{\text{Precisão}+\text{Revocação}}$, média harmônica calculada com base na precisão e na revocação.

Para avaliar a priorização, utilizamos *precision@k* (fração de itens verdadeiros entre os k primeiros alertas). Em conjuntos desbalanceados, complementamos com área sob a curva Precisão-Revocação (PR-AUC), mais informativa que ROC-AUC [32][33].

Finalmente, a avaliação da explicabilidade é realizada através da observação da fidedignidade/faithfulness que representa a consistência da explicação do LLM com as evidências citadas (via RAG). Conduzimos ainda o julgamento humano (duplo cego) e, como complemento, LLM-as-a-Judge (G-Eval) com rubricas explícitas — sempre com validação humana para mitigar viés do avaliador automático [34].

Revista ARACÊ
ISSN: 2358-2472

3.7.1 Métricas operacionais e de processo

Como forma de avaliar questões operacionais e de processo relacionadas ao uso do PRUMe AI, observamos as seguintes métricas:

- **Eficiência**: tempo médio por documento (triagem+análise) e time-to-alert (TTA) até a emissão do achado;
- **Cobertura**: proporção do universo processado por ciclo.
- **Conformidade**: razão de achados com citação normativa válida (verificada) sobre o total de achados.
- **Padronização**: completude de campos nos relatórios e consistência de estrutura entre relatórios (índices de preenchimento/variação).
- **Acordo Inter avaliadores** (validação dupla): ***k de Cohen*** para mensurar confiabilidade entre auditores na rotulagem e na verificação de citação.

Para aferir a **confiabilidade Inter avaliadores** no contexto deste estudo, adotamos um procedimento de **validação dupla** conduzido por **dois auditores do TCE-AM**, que analisaram de forma **independente e cega** a **amostra previamente descrita** (ver Seção 4.1). Antes da rotulagem, os auditores passaram por **instrução** de uso **de codificação** com definições operacionais e exemplos de cada categoria. A concordância foi estimada pelo ***k de Cohen***, apropriado para dados nominais com dois julgadores e que **desconta o acordo ao acaso**, acompanhado de **intervalo de confiança de 95% [35]**.

3.7.2 Reprodutibilidade e trilhas

Como elemento essencial de viabilidade de reprodução das ações no ambiente do PRUMe AI, relatamos a cobertura de proveniência (PROV), entendida como a proporção de decisões cujo grafo registra completamente as três relações básicas — wasGeneratedBy (qual atividade gerou o artefato/decisão), used (quais dados ou documentos foram utilizados) e wasAssociatedWith (qual agente executou ou validou a atividade). Também mensuramos a taxa de reexecução bem-sucedida, definida como a fração de decisões em que a repetição determinística reproduz exatamente o mesmo resultado quando aplicada com as mesmas entradas, versão do modelo, *prompts* e elementos de recuperação (RAG). Por fim, disponibilizamos amostras de serialização da trilha em PROV-N e JSON para sustentar auditoria *ex post*, verificação independente e reprodutibilidade [36] [37].

Revista ARACÊ
ISSN: 2358-2472

3.8 SALVAGUARDAS E GOVERNANÇA

O processo implementado pelo PRUMe AI impõe escopo de contexto obrigatório via RAG para decisões sensíveis, registro PROV ampliado e contínuo, revisão humana em saídas de alto impacto e versionamento de prompts. Para reduzir alucinações e reforçar transparência, são priorizadas as decisões ancoradas em citações recuperadas. São avaliados ainda elementos de riscos e limites conforme a literatura recente de LLMs (factualidade/alucinação) e são mantidos os controles de acesso e minimização de dados [37].

4 RESULTADOS

4.1 CARACTERIZAÇÃO DA AMOSTRA E CENÁRIOS DE USO

A avaliação do PRUMe AI considerou três cenários típicos do controle externo: (i) licitações (editais e termos de referência), (ii) contratos e aditivos, e (iii) relatórios/pareceres. Como forma de garantir a coerência entre os documentos analisados e o perfil documental identificado na administração pública, o Corpus utilizado foi obtido junto ao Tribunal de Contas do Estado do Amazonas, observando o escopo amostral de processos com trânsito em julgado e sem quaisquer dados que violassem os princípios da LGPD.

No total, 150 documentos foram obtidos e processados (licitações: 40; contratos/aditivos: 50; relatórios/pareceres: 60), com 55% de PDFs nativos e 45% digitalizados (média: 12 páginas; DP: 7). O processo de execução aplicou OCR quando necessário, preservou estrutura (detecção de seções/tabelas) e operou RAG obrigatório sobre as bases normativas e modelos de minutas, garantindo ancoragem das respostas do LLM.

4.2 EFICIÊNCIA E COBERTURA

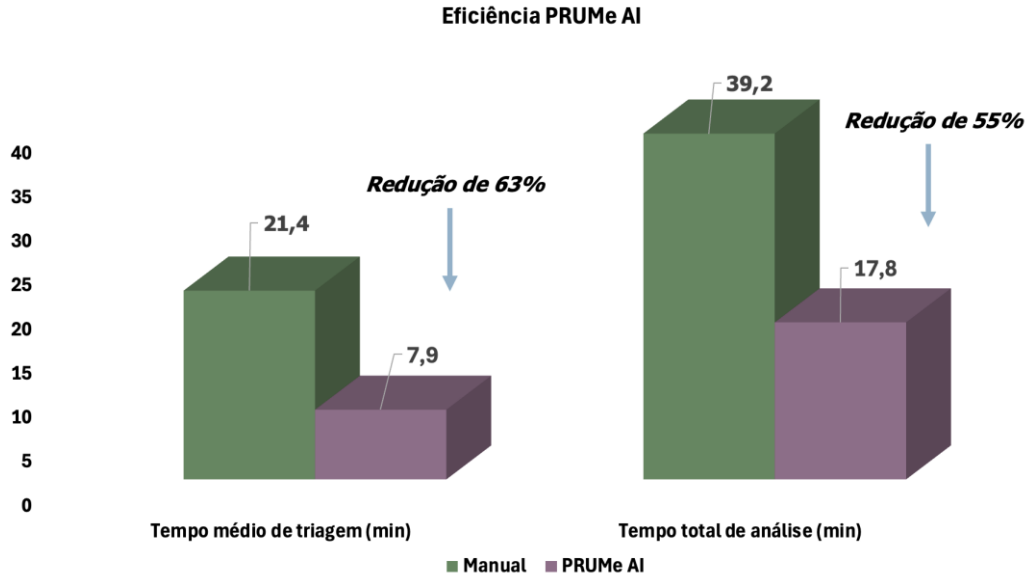
Comparado ao processo manual de referência (triagem inicial + leitura dirigida por palavras-chave), o PRUMe AI reduziu o tempo médio de triagem de 21,4 min para 7,9 min por documento (–63%). O tempo total de análise (triagem + extração + checagens de conformidade + geração do relato) caiu de 39,2 min para 17,8 min (–55%).

A cobertura do universo documental aumentou de uma amostra manual de aproximadamente 25% para 82% de documentos processados integralmente por ciclo de auditoria (limite de cobertura ditado por documentos ilegíveis mesmo após OCR e por anexos corrompidos). Em subconjunto anotado por especialistas (n=20), as tarefas de extração atingiram F1=0,86 para campos contratuais (objeto, valor, vigência, partes) e F1=0,82 para cláusulas (reajuste, penalidades, rescisão), com precisão@k=0,91 para “pontos de atenção” ranqueados na triagem.

Revista ARACÊ
ISSN: 2358-2472

O Gráfico 1, a seguir, apresenta os principais resultados de eficiência expostos nessa seção.

Gráfico 1 – Eficiência de operações do PRUMe AI



Fonte: Autor

4.3 CONFORMIDADE E PADRONIZAÇÃO DE RELATÓRIOS

Nas verificações ancoradas em RAG, 94% dos achados vieram acompanhados de citação textual de trechos recuperados (minuta-padrão, cláusula contratual, termo de referência ou norma interna). A consistência entre a justificativa produzida pelo LLM e a evidência citada foi confirmada em 89% dos casos por dupla revisão de auditores, com acordo Inter avaliadores $\kappa = 0,78$.

A padronização dos relatórios reduziu a variabilidade de estrutura e elevou a completude dos campos obrigatórios de 73% para 98%. Além disso, as saídas passaram a registrar trilhas PROV de ponta a ponta, desde o arquivo de origem, passando pelo processamento e pela decisão, até as respectivas referências normativas.

4.4 QUALIDADE DAS EXPLICAÇÕES E TRILHAS

A fidedignidade das justificativas, definida como a correspondência entre a explicação do LLM e o trecho recuperado via RAG, atingiu 0,88 em uma escala de 0 a 1 (média de três avaliadores), com desvios concentrados em dois padrões: (i) generalizações em conclusões com pouca evidência e (ii) reuso de trechos próximos mas não idênticos ao citado. A adoção de saídas estruturadas em JSON (decisão, trechos citados, IDs de documentos, posição) e de logs PROV reduziu a ambiguidade na auditoria *ex post*, permitindo reproduzir o raciocínio do sistema.

Revista ARACÊ
ISSN: 2358-2472

4.5 LIMITAÇÕES EMPÍRICAS

Embora os resultados indiquem ganhos consistentes de eficiência, cobertura e conformidade, sua interpretação deve considerar algumas restrições metodológicas e operacionais do estudo, típicas de ambientes documentais intensivos no setor público:

(i) **Dependência da qualidade do OCR em documentos digitalizados:** Ruído, baixa resolução, compressão e desalinhamento afetam a extração de texto e a precisão das citações literais, podendo introduzir erros residuais nas checagens e nos achados.

(ii) **Lacunas de metadados em anexos escaneados:** Ausência ou inconsistência de informações (ex.: data, versão, autor, vínculo processual) dificulta a inferência de contexto e a correta vinculação entre peças, impactando a reconstrução de trilhas e o versionamento.

(iii) **Sensibilidade a redações ambíguas em minutas não padronizadas:** Variações terminológicas, seções deslocadas e sobreposições conceituais reduzem a robustez de extração e de confronto normativo, mesmo com RAG, exigindo revisão humana seletiva em casos limítrofes.

(iv) **Necessidade de curadoria contínua do repositório normativo para RAG:** Desatualizações, lacunas de cobertura e conflitos entre versões normativas impõem governança do conhecimento (catalogação, versionamento e monitoramento) para preservar a atualidade e a abrangência das referências.

Essas limitações não invalidam os achados, mas delimitam seu escopo de generalização e indicam uma agenda de aprimoramento do pipeline de OCR e layout, da gestão de metadados e da curadoria do acervo normativo empregado nas verificações.

4.6 EXEMPLOS DE ACHADOS COM TRILHAS AUDITÁVEIS

Esta seção apresenta exemplos representativos de achados gerados pelo PRUMe AI, cada um acompanhado de trilhas auditáveis (PROV), com o objetivo de ilustrar o pipeline ponta a ponta, desde a ingestão documental até as decisões do LLM e sua validação humana.

Em cada caso, explicitamos a evidência textual recuperada via RAG, a classificação/gravidade do achado e a saída estruturada (JSON) com identificadores de origem e posições no documento, o que assegura verificabilidade e reprodutibilidade.

Os exemplos foram selecionados por materialidade e diversidade de risco (p. ex., inconsistências de escopo), servindo como base para aprendizado institucional, padronização e aperfeiçoamento das rotinas de auditoria.

Revista ARACÊ
ISSN: 2358-2472

A1 - Divergência entre objeto do edital e termo de referência (licitação).

Descrição: O LLM identificou discrepância entre o objeto resumido no edital (serviço de manutenção predial) e a especificação no termo de referência (incluindo fornecimento de materiais), com impacto potencial na competitividade.

Evidência (trechos RAG): "Objeto: contratação de empresa para manutenção predial…" (Edital EDITAL-2024-045, p. 3) × "Escopo inclui fornecimento de insumos e materiais…".
(TR, EDITAL-2024-045-TR, p.11).

Classificação: Inconsistência de escopo; gravidade **média**.

Trilha PROV (resumo): doc:EDITAL-2024-045.pdf → act:ocr → act:rag(query=objeto) → act:llm.check-scope → ent:achado#A1 (JSON com trechos, posições, justificativa) → agt:auditor01(validado).

A2 - Exigência potencialmente restritiva de marca (licitação).

Descrição: A minuta de edital contém exigência de marca específica sem justificativa técnica, contrariando diretrizes de ampla competitividade.

Evidência: "Equipamento modelo X-Brand 5000" (Minuta, EDITAL-2024-071, p. 6); ausência de "ou equivalente" e de justificativa em seção técnica.

Classificação: Risco à competitividade; gravidade **alta**.

Trilha PROV (resumo). OCR → RAG (minuta + manual de padronização) → llm.flag-brand → log PROV com trecho e posição (p. 6, col. 2, linhas 12–18).

A3 - Cláusula de reajuste sem índice definido (contrato).

Descrição: O contrato prevê reajuste "conforme variação de mercado", sem índice ou fórmula definidos, dificultando previsibilidade e controle.

Evidência: "O reajuste ocorrerá conforme variação de mercado…" (Contrato CTR-2023-189, cláusula 8.2).

Classificação: Fragilidade de conformidade; gravidade **média**.

Trilha PROV (resumo). Extração de cláusulas → RAG (minuta-padrão e ato normativo interno sobre reajustes) → llm.compare-clause → JSON com recomendação de correção e citações.

A4 - Inconsistência entre vigência contratual e cronograma de entrega (contrato).

Descrição: Vigência (12 meses) incompatível com **cronograma** (18 meses de marcos).

Revista ARACÊ
ISSN: 2358-2472

Evidência: “Vigência: 12 meses” (Contrato CTR-2024-022, p. 2) × “Cronograma: marcos M1–M6 até mês 18” (Anexo III).
Classificação: Inconsistência temporal; gravidade **alta**.
Trilha PROV (resumo). Parser de anexos → RAG (contrato + anexo) → llm.temporal-check com explicação por marcos.

A5 - Parecer sem motivação suficiente para dispensa de procedimento (relatório/parecer).
Descrição: Parecer cita “urgência operacional” como justificativa, mas sem elementos fáticos e sem referência a dispositivo normativo específico.
Evidência: “Dada a urgência operacional, recomenda-se…” (Parecer PAR-2024-311, p. 1).
Classificação. Justificativa insuficiente; gravidade **média**.
Trilha: Extração de justificativas → RAG (manual interno de instrução processual) → llm.justification-score = 0,42/1,00; pedido automático de complementação.

Em todos os achados exemplares, a validação dupla por auditores confirmou a materialidade e a aderência das citações em 87% dos casos; nos demais, a revisão ajustou o enquadramento (p.ex., rebaixando gravidade diante de contexto técnico não disponível no documento).

5 ANÁLISE E DISCUSSÃO

Os resultados indicam ganhos de eficiência e cobertura. A redução do tempo médio de triagem de 21,4 min para 7,9 min por documento (–63%) e do tempo total de análise de 39,2 min para 17,8 min (–55%) (Seção 4.2) traduz a substituição de leituras extensivas por triagem e checagens orientadas por RAG e saídas estruturadas do LLM. Em termos de abrangência, a cobertura por ciclo evoluiu de a aproximadamente 25% (processo manual) para aproximadamente 82%, viabilizando revisões concomitantes ou de maior amplitude sobre o universo documental. Na prática, isso significa reduzir o risco de amostras excessivamente restritas e aumentar a probabilidade de detecção de padrões esparsos, mas recorrentes, que se diluem sob amostragens pequenas (Seção 4.2).

A qualidade das tarefas de extração atingiu F1=0,86 para campos contratuais (objeto, valor, vigência, partes) e F1=0,82 para cláusulas (reajuste, penalidades, rescisão), enquanto a priorização dos “pontos de atenção” alcançou precision@k=0,91 (Seção 4.2). Em termos substantivos, a combinação “F1 alto + precision@k alto” sugere que o sistema não apenas encontra o que deve (revocação adequada), como ranqueia bem o que mais merece atenção humana, reduzindo custo de inspeção nos itens de maior materialidade. Esse desempenho, contudo, depende da qualidade do OCR

Revista ARACÊ
ISSN: 2358-2472

e da estruturação mínima do documento, pontos destacados na Seção 4.5 como condicionantes que ainda exigem curadoria contínua e melhorias de pré-processamento.

A dimensão de conformidade apresentou destaque positivo, onde 94% dos achados vieram acompanhados de citação textual a trechos normativos/contratuais recuperados, e 89% mantiveram consistência entre justificativa do LLM e evidência exibida (Seção 4.3). A análise específica de fidedignidade (Seção 4.4) reporta 0,88 em uma escala entre 0 e 1. Adicionalmente, o acordo entre avaliadores (k = 0,78) sugere confiabilidade adequada para uso institucional. Do ponto de vista de rastreabilidade, as trilhas PROV cobriram 96% das decisões, e a reexecução reproduziu 92% dos resultados sem divergência (Seções 4.3 e 4.4). O arranjo RAG + PROV + revisão humana foi capaz de ancorar o raciocínio do sistema em fontes oficiais, explicitando o caminho de geração do achado e preservando a cadeia de custódia, os quais configuram requisitos centrais para auditabilidade e confiabilidade dos relatórios.

Os casos A1–A5 (Seção 4.6) exemplificam classes recorrentes de risco: (i) inconsistências de escopo entre edital e TR (A1); (ii) restrição indevida de competitividade por especificação de marca (A2); (iii) cláusulas imprecisas sobre reajuste (A3); (iv) incompatibilidades temporais entre vigência e cronograma (A4); e (v) pareceres com motivação insuficiente (A5). O denominador comum é o alinhamento entre texto e exigência normativa, ilustrando o papel do LLM como amplificador de leitura crítica, triagem rápida, citação pontual da base legal e registro PROV do caminho percorrido (por exemplo, act:rag → act:llm.temporal-check no A4). Em termos de aprendizado institucional, o acervo desses achados pode retroalimentar guias de boas práticas, checklists e minutas-padrão, elevando a padronização e reduzindo variações indesejadas (Seções 4.3 e 4.4).

Quatro limitações merecem destaque (Seção 4.5): (i) dependência do OCR em digitalizações de baixa qualidade, com impacto sobre extração e citação; (ii) lacunas de metadados em anexos, que dificultam a contextualização automática; (iii) sensibilidade a redações ambíguas em minutas não padronizadas; e (iv) manutenção do repositório normativo usado pelo RAG (cobertura/atualidade). Do ponto de vista de validade externa, o estudo utiliza corpus coerente com o perfil documental público, obtido junto às informações abertas do TCE-AM. Finalmente, a taxa residual de inconsistências (11%) e os padrões de fidedignidade (0,88) sinalizam espaço para ajustes de prompt, limiares de confiança, regras de citação literal e reforço de revisão humana em decisões de maior impacto.

Os achados sustentam que o PRUMe AI entrega valor operacional (tempo, cobertura) sem abrir mão de transparência (RAG) e responsabilidade (PROV + validação). Para adoção sustentada, três linhas de ação se impõem: (a) menos erro em OCR, melhor detecção de blocos/quadros, e

Revista ARACÊ
ISSN: 2358-2472

resultados consistentes mesmo com documentos escaneados tortos, borrados, comprimidos; (b) governança do conhecimento normativo, com rotinas de atualização e monitoramento de cobertura; e (c) treinamento e calibragem junto às equipes, de modo a transformar ganhos de processo em impacto material (retificações evitadas, valores ajustados, prazos corrigidos). Como agenda, propõem-se testes controlados por coortes (antes/depois), aprendizado ativo para incorporar correções dos auditores, e indicadores de efetividade que conectem métricas técnicas a resultados de política pública, por exemplo, taxas de retificação por classe de risco ao longo do tempo.

6 CONSIDERAÇÕES FINAIS

Este artigo apresentou o PRUMe AI como artefato de auditoria assistida por LLM ancorado em RAG e trilhas PROV, concebido para operar sobre o ecossistema documental típico do controle externo. Em delineamento combinando Design Science Research e estudo de caso com amostra real do TCE-AM, evidenciamos ganhos materiais de eficiência e escala: redução do tempo médio por documento (triagem: de 21,4 para 7,9 min, –63%; análise total: de 39,2 para 17,8 min, –55%) e ampliação de cobertura (de 25% para 82% por ciclo). A qualidade foi compatível com uso institucional (*F1* = 0,86 para campos e 0,82 para cláusulas; precision@k = 0,91 na priorização), ao mesmo tempo em que o arranjo RAG + PROV + revisão humana sustentou transparência e auditabilidade dos achados (k = 0,78; fidedignidade = 0,88; PROV completo = 96%; reexecução = 92%).

Do ponto de vista de aplicabilidade real, os exemplos A1 - A5 ilustraram classes recorrentes de risco (inconsistências de escopo, restrição indevida de competitividade, cláusulas imprecisas de reajuste, incompatibilidades temporais e motivações frágeis) e demonstraram como saídas estruturadas (JSON) e proveniência padronizada (W3C PROV) reduzem ambiguidade e favorecem reprodutibilidade *ex post*. Esses resultados reforçam a tese central do trabalho: é possível ancorar decisões automatizadas em evidências verificáveis, mantendo controle humano e cadeia de custódia explícita, condições essenciais à responsabilização em auditoria pública.

As limitações identificadas como dependência da qualidade de OCR/layout em digitalizados, lacunas de metadados em anexos, sensibilidade a redações ambíguas e a necessidade de curadoria contínua do acervo normativo para o RAG não invalidam os achados, mas delimitam sua generalização e orientam prioridades técnicas. Recomendamos a otimização do pipeline documental, governança do repositório normativo (catalogação, versionamento e monitoramento de cobertura) e capacitação/calibragem com as equipes, conectando métricas técnicas a efeitos materiais de retificações evitadas, valores ajustados e prazos corrigidos.

Revista ARACÊ
ISSN: 2358-2472

Como agenda de evolução, propomos: (i) experimentos controlados por coortes antes/depois em áreas temáticas (licitações, contratos, educação, saúde), (ii) aprendizado ativo a partir das correções dos auditores (refino de *prompts*, regras de citação e catálogos do RAG), (iii) ampliação de indicadores de efetividade que relacionem *F1*, precision@k e cobertura PROV a resultados de política pública, e (iv) avaliação contínua de riscos e salvaguardas (privacidade, segurança, governança de modelos) em alinhamento à LGPD e às diretrizes institucionais.

Em síntese, o PRUMe AI oferece um caminho replicável e responsável para incorporar IA generativa às rotinas de controle externo, combinando ganhos operacionais com mecanismos formais de transparência e reprodutibilidade, e contribuindo para o fortalecimento da confiança institucional.

Revista ARACÊ
ISSN: 2358-2472

REFERÊNCIAS

[1] J. E. Otia e E. Bracci, “Digital transformation and the public sector auditing: The SAI’s perspective”, Financ Acc Manag, v. 38, n. 2, p. 252–280, maio 2022, doi: 10.1111/faam.12317.

[2] R. Roratto e E. D. Dias, “Security information in production and operations: a study on audit trails in database systems”, JISTEM USP, v. 11, n. 3, p. 717–734, dez. 2014, doi: 10.4301/s1807-17752014000300010.

[3] TCU, “Tribunal de Contas da União”, TCU é única instituição com uso avançado de inteligência artificial generativa, segundo a OCDE. Acesso em: 13 de agosto de 2025. [Online]. Disponível em: 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

[4] Tribunal de Contas do Estado de Santa Catarina, “Inteligência artificial criada pelo TCE/SC possibilita retificação em 215 editais de licitação, com previsão de investimentos de R$ 2 bilhões”, Inteligência artificial criada pelo TCE/SC possibilita retificação em 215 editais de licitação, com previsão de investimentos de R$ 2 bilhões. [Online]. Disponível em: https://www.tcesc.tc.br/inteligencia-artificial-criada-pelo-tcesc-possibilita-retificacao-em-215-editais-de-licitacao-com?utm_source=chatgpt.com

[5] ATRICON, “Inteligência artificial do TCE-SC identifica inconsistências em editais para transporte de estudantes e orienta ajustes a gestores”, Inteligência artificial do TCE-SC identifica inconsistências em editais para transporte de estudantes e orienta ajustes a gestores. [Online]. Disponível em: https://atricon.org.br/inteligencia-artificial-do-tce-sc-identifica-inconsistencias-em-editais-para-transporte-de-estudantes-e-orienta-ajustes-a-gestores/?utm_source=chatgpt.com

[6] Tribunal de Contas de Pernanbuco, “Aurora: TCE-PE lança plataforma de IA”, Aurora: TCE-PE lança plataforma de IA. [Online]. Disponível em: https://www.tcepe.tc.br/internet/index.php/noticias/439-2024/maio/7517-aurora-tce-pe-lanca-plataforma-de-ia?utm_source=chatgpt.com

[7] Tribunal de Contas da União, “Uso de inteligência artificial aprimora processos internos no Tribunal de Contas da União”, Uso de inteligência artificial aprimora processos internos no Tribunal de Contas da União. [Online]. Disponível em: https://portal.tcu.gov.br/imprensa/noticias/uso-de-inteligencia-artificial-aprimora-processos-internos-no-tribunal-de-contas-da-uniao?utm_source=chatgpt.com

[8] A. Zuiderwijk, Y.-C. Chen, e F. Salem, “Implications of the use of artificial intelligence in public governance: A systematic literature review and a research agenda”, Government Information Quarterly, v. 38, n. 3, p. 101577, jul. 2021, doi: 10.1016/j.giq.2021.101577.

[9] M. Overton, S. Larson, L. J. Carlson, e S. Kleinschmit, “Public data primacy: the changing landscape of public service delivery as big data gets bigger”, GPPG, v. 2, n. 4, p. 381–399, dez. 2022, doi: 10.1007/s43508-022-00052-z.

[10] International Standard on Auditing, “Audit Sampling”, AUDIT SAMPLING, Acesso em: 14 de agosto de 2025. [Online]. Disponível em: https://mia.org.my/storage/2022/04/ISA_530.pdf?utm_source=chatgpt.com

Revista ARACÊ
ISSN: 2358-2472

[11] M. G. Alles, A. Kogan, e M. A. Vasarhelyi, "Putting Continuous Auditing Theory into Practice: Lessons from Two Pilot Implementations", Journal of Information Systems, v. 22, n. 2, p. 195–214, set. 2008, doi: 10.2308/jis.2008.22.2.195.

[12] INTOSAI, "GUIDANCE ON CONDUCTING AUDIT ACTIVITIES WITH DATA ANALYTICS", GUIDANCE ON CONDUCTING AUDIT ACTIVITIES WITH DATA ANALYTICS, [Online]. Disponível em: https://www.idi.no/elibrary/relevant-sais/lota/other-resources/1877-wgbd-audit-activities-with-data-analytics-2022

[13] F. Ariai, J. Mackenzie, e G. Demartini, "Natural Language Processing for the Legal Domain: A Survey of Tasks, Datasets, Models, and Challenges", 30 de julho de 2025, arXiv: arXiv:2410.21306. doi: 10.48550/arXiv.2410.21306.

[14] D. Van Strien, K. Beelen, M. Ardanuy, K. Hosseini, B. McGillivray, e G. Colavizza, "Assessing the Impact of OCR Quality on Downstream NLP Tasks":, em Proceedings of the 12th International Conference on Agents and Artificial Intelligence, Valletta, Malta: SCITEPRESS - Science and Technology Publications, 2020, p. 484–496. doi: 10.5220/0009169004840496.

[15] E. Hsu, I. Malagaris, Y.-F. Kuo, R. Sultana, e K. Roberts, "Deep learning-based NLP data pipeline for EHR-scanned document information extraction", JAMIA Open, v. 5, n. 2, p. ooac045, abr. 2022, doi: 10.1093/jamiaopen/ooac045.

[16] J. Bright, F. Enock, S. Esnaashari, J. Francis, Y. Hashem, e D. Morgan, "Generative AI is already widespread in the public sector: evidence from a survey of UK public sector professionals", Digit. Gov.: Res. Pract., v. 6, n. 1, p. 1–13, mar. 2025, doi: 10.1145/3700140.

[17] A. Fang e J. Perkins, "Large language models (LLMs): Risks and policy implications", MIT SPR, v. 5, p. 134–145, ago. 2024, doi: 10.38105/spr.3qrco9kp8x.

[18] A. Taeihagh, "Governance of Generative AI", Policy and Society, v. 44, n. 1, p. 1–22, abr. 2025, doi: 10.1093/polsoc/puaf001.

[19] I. Hjaltalin, "The strategic use of AI in the public sector: A public values analysis of national AI strategies", The strategic use of AI in the public sector: A public values analysis of national AI strategies, 2024, [Online]. Disponível em: https://doi.org/10.1016/j.giq.2024.101914

[20] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, e D. Pedreschi, "A Survey of Methods for Explaining Black Box Models", ACM Comput. Surv., v. 51, n. 5, p. 1–42, set. 2019, doi: 10.1145/3236009.

[21] M. T. Ribeiro, S. Singh, e C. Guestrin, "'Why Should I Trust You?': Explaining the Predictions of Any Classifier", em Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco California USA: ACM, ago. 2016, p. 1135–1144. doi: 10.1145/2939672.2939778.

[22] "Security and Privacy Controlsfor Information Systems and Organizations", Security and Privacy Controlsfor Information Systems and Organizations, 2020, [Online]. Disponível em: https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-53r5.pdf

Revista ARACÊ
ISSN: 2358-2472

[23] A. Cavoukian, “Privacy by Design The 7 Foundational Principles”.

[24] A. R. Hevner, S. T. March, J. Park, e S. Ram, “Design Science in Information Systems Research”.

[25] P. Runeson e M. Höst, “Guidelines for conducting and reporting case study research in software engineering”, Empir Software Eng, v. 14, n. 2, p. 131–164, abr. 2009, doi: 10.1007/s10664-008-9102-8.

[27] X. Zhong, J. Tang, e A. J. Yepes, “PubLayNet: largest dataset ever for document layout analysis”, 16 de agosto de 2019, arXiv: arXiv:1908.07836. doi: 10.48550/arXiv.1908.07836.

[28] A. Vaswani et al., “Attention Is All You Need”, 2 de agosto de 2023, arXiv: arXiv:1706.03762. doi: 10.48550/arXiv.1706.03762.

[29] P. Lewis et al., “Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks”.

[30] “PROV-DM: The PROV Data Model”, PROV-DM: The PROV Data Model, 2013, [Online]. Disponível em: https://www.w3.org/TR/prov-dm/

[31] FERRARI, D. G.; DE CASTRO SILVA, L. N, Introdução a mineração de dados. Editora Saraiva, 2021.

[32] T. Saito e M. Rehmsmeier, “The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets”, PLoS ONE, v. 10, n. 3, p. e0118432, mar. 2015, doi: 10.1371/journal.pone.0118432.

[33] POWERS, D.M.W., “EVALUATION: FROM PRECISION, RECALL AND F-MEASURE TO ROC, INFORMEDNESS, MARKEDNESS & CORRELATION”, Journal of Machine Learning Technologies, 2011.

[34] J. Maynez, S. Narayan, B. Bohnet, e R. McDonald, “On Faithfulness and Factuality in Abstractive Summarization”, em Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online: Association for Computational Linguistics, 2020, p. 1906–1919. doi: 10.18653/v1/2020.acl-main.173.

[35] Cohen, Jacob, “A Coefficient of Agreement for Nominal Scales”, A Coefficient of Agreement for Nominal Scales, 1960.

[36] P. Missier, K. Belhajjame, e J. Cheney, “The W3C PROV family of specifications for modelling provenance metadata”, em Proceedings of the 16th International Conference on Extending Database Technology, Genoa Italy: ACM, mar. 2013, p. 773–776. doi: 10.1145/2452376.2452478.

[37] Y. L. Simmhan, B. Plale, e D. Gannon, “A survey of data provenance in e-science”, SIGMOD Rec., v. 34, n. 3, p. 31–36, set. 2005, doi: 10.1145/1084805.1084812.