UNIVERSIDADE FEDERAL DE SANTA CATARINA
CENTRO SOCIOECONÔMICO
DEPARTAMENTO DE CIÊNCIAS CONTÁBEIS
CURSO DE CIÊNCIAS CONTÁBEIS

Daniel Piotroski da Silva

EXTRAÇÃO DE INFORMAÇÃO EM CONTRATOS PÚBLICOS: AVALIAÇÃO DE MODELO DE LINGUAGEM DE LARGA ESCALA

Florianópolis
2025

Daniel Piotroski da Silva

EXTRAÇÃO DE INFORMAÇÃO EM CONTRATOS PÚBLICOS: avaliação de modelo de linguagem de larga escala

Trabalho de Conclusão de Curso submetido ao curso de Ciências Contábeis do Centro Socioeconômico da Universidade Federal de Santa Catarina como requisito parcial para a obtenção do título de Bacharel em Ciências Contábeis.

Orientador: Prof. Dr. Marcelo Machado de Freitas

Florianópolis

2025

Ficha catalográfica gerada por meio de sistema automatizado gerenciado pela BU/UFSC.
Dados inseridos pelo próprio autor.

da Silva, Daniel Pìotroski
EXTRAÇÃO DE INFORMAÇÃO EM CONTRATOS PÚBLICOS: : avaliação de modelo de linguagem de larga escala / Daniel Pìotroski da Silva ; orientador, Marcelo Machado de Freitas, 2025.
63 p.

Trabalho de Conclusão de Curso (graduação) - Universidade Federal de Santa Catarina, Centro Socioeconômico, Graduação em Ciências Contábeis, Florianópolis, 2025.

Inclui referências.

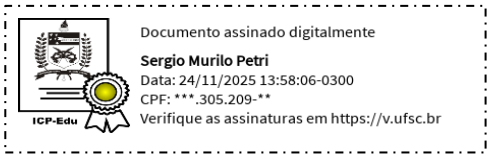
1. Ciências Contábeis. 2. Auditoria. 3. Inteligência Artificial. 4. Extração de Dados. I. Freitas, Marcelo Machado de. II. Universidade Federal de Santa Catarina. Graduação em Ciências Contábeis. III. Título.

Daniel Piotroski da Silva

EXTRAÇÃO DE INFORMAÇÃO EM CONTRATOS PÚBLICOS: avaliação de modelo de linguagem de larga escala

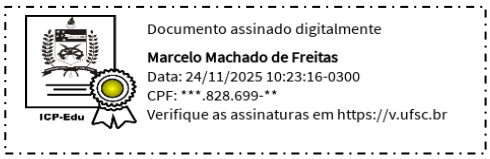
Este Trabalho de Conclusão de Curso foi julgado adequado para obtenção do título de Bacharel e aprovado em sua forma final pelo Curso de Ciências Contábeis.

Florianópolis, 23 de novembro de 2025.



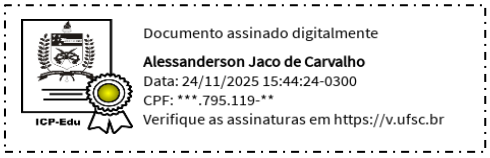
Coordenação do Curso

Banca examinadora



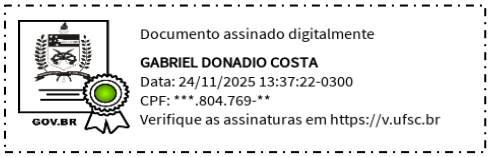
Prof. Marcelo Machado de Freitas, Dr.

Orientador



Prof. Alessanderson Jaco de Carvalho, Dr.

Universidade Federal de Santa Catarina - UFSC



Prof. Gabriel Donadio Costa, Me.

Universidade Federal de Santa Catarina - UFSC

Florianópolis, 2025.

AGRADECIMENTOS

Agradeço, primeiramente, ao meu professor orientador, Dr. Marcelo Machado de Freitas, pela dedicação, disponibilidade e incentivo constantes ao longo de todo o desenvolvimento deste trabalho. Estendo meus agradecimentos aos demais professores que, direta ou indiretamente, me forneceram subsídios valiosos que me conduziram ao limiar desta investigação. Por fim, agradeço também aos amigos e colegas pelas frequentes trocas de conhecimentos, experiências e encorajamento.

"artificial intelligence, with its immense potential, which nevertheless requires responsibility and discernment in order to ensure that it can be used for the good of all, so that it can benefit all of humanity. This responsibility concerns everyone in proportion to his or her age and role in society."

(Papa Leão XIV, 2025)

RESUMO

Esta pesquisa tem por objetivo avaliar o grau de êxito de um modelo de linguagem de larga escala na extração de informações em contratos públicos, visando analisar a capacidade desses modelos de interpretar, organizar e estruturar dados complexos provenientes de documentos administrativos. O trabalho aplica uma metodologia descritiva com abordagem quali-quantitativa. A amostra compreende 100 contratos do Departamento de Logística em Saúde (DLOG) do Ministério da Saúde, referentes ao ano de 2024. Utilizou-se o modelo Gemini 2.5 Pro, guiado por um prompt estruturado, para extrair nove variáveis predefinidas (como número, data, objeto, fornecedor, CNPJ e valor). Paralelamente, uma análise qualitativa mensurou a qualidade textual (OCR) dos arquivos. Os resultados quantitativos indicam um êxito excepcional: oito dos nove campos alcançaram 100% de acurácia, com o campo textual 'Objeto' atingindo 91,13% de média. A eficiência também se mostrou elevada, com um tempo médio de processamento de 19,53 segundos por contrato. A principal conclusão é a resiliência do modelo, cuja performance de extração demonstrou ser largamente independente da qualidade textual (OCR) dos documentos. O estudo evidencia que a tecnologia está madura para superar a análise por amostragem, viabilizando a análise censitária.

Palavras-chave: Modelos de Linguagem de Larga Escala; Extração de Dados; Auditoria.

ABSTRACT

This research aims to evaluate the degree of success of a large-scale language model in extracting information from public contracts, with a view to analyzing the ability of these models to interpret, organize, and structure complex data from administrative documents. The study applies a descriptive methodology with a qualitative-quantitative approach. The sample comprises 100 contracts from the Department of Health Logistics (DLOG) of the Ministry of Health, referring to the year 2024. The Gemini 2.5 Pro model, guided by a structured prompt, was used to extract nine predefined variables (such as number, date, object, supplier, CNPJ, and value). At the same time, a qualitative analysis measured the textual quality (OCR) of the files. The quantitative results indicate exceptional success: eight of the nine fields achieved 100% accuracy, with the textual field 'Object' reaching an average of 91.13%. Efficiency was also high, with an average processing time of 19.53 seconds per contract. The main conclusion is the resilience of the model, whose extraction performance proved to be largely independent of the textual quality (OCR) of the documents. The study shows that the technology is mature enough to overcome sampling analysis, enabling census analysis.

Keywords: Large Language Models; Data Extraction; Auditing.

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

O uso da inteligência artificial no cenário contemporâneo tem se tornado cada vez mais frequente e nos mais variados contextos concebíveis, inclusive por auditores ao redor do mundo todo. Essa transformação na área da auditoria foi documentada por Kokina e Davenport (2017), que destacam como as grandes firmas globais do setor já utilizam IA para automatizar tarefas repetitivas, analisar contratos e extrair dados relevantes, e avaliar riscos com maior precisão. Dentro do rol de possibilidades que são ofertadas pela tecnologia, existem os Large Language Models (LLMs) (tradução própria: Modelos de linguagem de larga escala), Zhao et al. (2023) definem Large Language Models (LLMs) como modelos de linguagem capazes de compreender a linguagem natural e resolver tarefas complexas.

Seguindo a conceituação de Di Pietro (2025), os contratos públicos são os acordos que a Administração Pública estabelece com entidades privadas ou outros órgãos públicos para atender a uma finalidade de interesse coletivo. Eles se distinguem por serem governados por um conjunto de regras específicas do direito público. Diante da complexidade e do volume desses acordos, a sua gestão e fiscalização são um desafio constante, e é nesse contexto que o potencial de ferramentas para analisar grandes volumes de texto e identificar padrões complexos torna-se particularmente relevante para o setor público (Grimmer e Stewart, 2013).

Nos últimos anos, a área de Inteligência Artificial (IA) tem testemunhado, também, um grande crescimento no âmbito da pesquisa acadêmica, uma tendência quantificada pelo aumento no volume de publicações científicas e trabalhos em conferências sobre a temática, conforme documentado pelo AI Index Report 2025 (MASLEJ et al., 2025). Esse aumento na produção científica está associado ao desenvolvimento de algoritmos e modelos e à percepção da IA como uma tecnologia de amplo impacto.

Apesar dos avanços na digitalização e disponibilização de documentos públicos, conforme se vê pela promulgação do Decreto nº 10.278, de 18 de março de 2020, ainda persistem desafios relacionados à padronização e acessibilidade das informações neles contidas. Frequentemente, os dados são apresentados de forma desestruturada, com ausência de tabelas relacionais consistentes e formatos heterogêneos, o que dificulta a extração precisa de informações relevantes. Essas deficiências podem comprometer o controle social, a fiscalização por órgãos competentes e a garantia de direitos fundamentais relacionados ao acesso à informação (Bataglia & Farranha, 2018).

A Lei nº 12.527 de 18 de novembro de 2011, conhecida como Lei de Acesso à Informação (LAI), estabelece que a publicidade é a regra e o sigilo, a exceção, determinando

que os órgãos públicos devem divulgar, ativa ou passivamente, informações de interesse coletivo, de forma clara e acessível. No entanto, a falta de mecanismos eficientes para estruturar e verificar os dados agrava o risco de erros, omissões e interpretações equivocadas, o que poderia resultar em prejuízo à confiança da sociedade nas instituições (Janssen, Charalabidis e Zuiderwijk, 2012).

A abertura de dados governamentais é considerada um meio para promover a participação cidadã e a inovação na gestão pública. A disponibilização de dados de forma estruturada e acessível permite que cidadãos, pesquisadores e organizações da sociedade civil possam analisar, reutilizar e gerar valor a partir dessas informações (Coelho et al., 2018).

A apresentação desestruturada dos dados públicos, especialmente nos contratos administrativos, é capaz de obstruir a verificação e o cruzamento de informações relevantes. Em muitos casos, a ausência de tabelas relacionais idôneas impossibilita a comparabilidade dos documentos e pode dificultar a transparência exigida pela Lei de Acesso à Informação (Lei nº 12.527/2011). Assim, justifica-se a necessidade de investigar a verificabilidade dos dados disponíveis, buscando identificar o grau de organização das informações.

1.1 OBJETIVO GERAL

O objetivo geral deste trabalho é avaliar o grau de êxito de um modelo de linguagem de larga escala na extração de informações em contratos públicos, visando analisar a capacidade desses modelos de interpretar, organizar e estruturar dados complexos provenientes de documentos administrativos.

Busca-se, com isso, aferir a acurácia das soluções baseadas em IA para o tratamento de grandes volumes de dados públicos, em especial os referentes aos Contratos do Departamento de Logística em Saúde firmados pelo Ministério da Saúde do Brasil.

1.2 OBJETIVOS ESPECÍFICOS

Para alcançar o objetivo geral desta pesquisa, tem-se por objetivos específicos:

a) Identificar a verificabilidade de características e informações dispostas nos contratos públicos;

b) Mensurar a acurácia e a eficiência do modelo de linguagem de larga escala na extração automatizada de informações contratuais; e
c) Analisar a relação entre a qualidade documental dos contratos e o desempenho do modelo, verificando a influência de fatores textuais sobre a extração de dados.

1.3 JUSTIFICATIVA

Os modelos de linguagem de larga escala (Large Language Models – LLMs) têm sido estudados como uma possível solução nesse cenário, devido à sua capacidade de interpretar e organizar grandes volumes de dados textuais. Um estudo recente de Li et al. (2024) demonstrou que a aplicação de LLMs em auditorias públicas possibilitou um aumento notável da acurácia na verificação de informações (96% de precisão) e uma redução de 83% no tempo de análise, em comparação a métodos tradicionais. Ao realizar testes em bases públicas, o estudo evidenciou que essas tecnologias podem contribuir para a eficiência e a confiabilidade na análise de documentos governamentais. Assim, justifica-se a avaliação da efetividade dos LLMs especificamente no contexto dos contratos públicos, contribuindo para o desenvolvimento da pesquisa científica na área de ciência de dados aplicada à accountability.

A utilização de um modelo de linguagem para extrair e estruturar dados públicos aponta para aplicações práticas, como a otimização de relatórios e a melhoria do controle social (Schmidt, Maiden & Maiden, 2023). Embora a tecnologia possibilite uma análise mais precisa e acessível, sua abordagem também apresenta desafios, notadamente a necessidade de validação dos dados extraídos e a redução de eventuais vieses algorítmicos. Dessa forma, justifica-se uma análise crítica desta técnica, buscando identificar não só os benefícios, mas também as limitações que podem afetar a gestão pública e a fiscalização.

2 REFERENCIAL TEÓRICO

2.1 FUNDAMENTOS DA ADMINISTRAÇÃO PÚBLICA E CONTRATOS ADMINISTRATIVOS

A atuação da Administração Pública na gestão de contratos, firmados para atender ao interesse coletivo, é regida por um regime jurídico próprio e pautada por princípios constitucionais como legalidade, publicidade e eficiência (Di Pietro, 2025).

2.1.1 Conceito e regime jurídico dos contratos públicos

A atuação administrativa, especialmente em matéria contratual, exige a observância de um regime jurídico próprio, orientado pela supremacia do interesse público e pela legalidade estrita. Di Pietro (2025) conceitua os contratos administrativos como ajustes firmados entre a Administração Pública e particulares, pessoas físicas ou jurídicas, de direito público ou privado, com vistas à realização de objetivos de interesse coletivo.

Esse regime é regido por normas específicas, como a Lei de Licitações (Lei nº 14.133/2021), além de dispositivos da Lei de Acesso à Informação (Lei nº 12.527/2011), que impõem deveres de publicidade e transparência. Diferentemente dos contratos privados, os contratos administrativos incorporam cláusulas exorbitantes, que conferem à Administração Pública prerrogativas específicas, como a possibilidade de alteração ou rescisão unilateral, aplicação de sanções e fiscalização da execução contratual, visando resguardar o interesse público (Di Pietro, 2025).

2.1.2 Princípios constitucionais aplicáveis

A atuação da Administração Pública, inclusive na celebração e execução de contratos, deve respeitar os princípios expressos no caput do art. 37 da Constituição Federal de 1988: legalidade, impessoalidade, moralidade, publicidade e eficiência. Esses princípios norteiam

toda a atuação administrativa e são fundamentais para garantir a transparência e o controle externo e interno dos contratos públicos.

Além desses, aplicam-se também os princípios implícitos (Mello, 2023), como a supremacia do interesse público, continuidade do serviço público, e da vinculação ao instrumento convocatório, os quais impõem à Administração Pública o dever de garantir que os contratos estejam disponíveis, auditáveis e estruturados de modo a permitir seu acompanhamento por órgãos de controle e pela sociedade civil.

2.1.3 Fiscalização e controle dos contratos

Conforme preceitua Carvalho Filho (2025), o controle dos contratos administrativos se dá de forma interna (pelas próprias unidades administrativas) e externa (por tribunais de contas e órgãos de controle). Esse controle demanda acesso claro, preciso e estruturado às informações contratuais, razão pela qual a organização dos dados é um fator crucial para a efetividade da fiscalização.

Entretanto, observa-se que, em diversas situações, as informações contratuais podem ser apresentadas de forma fragmentada ou desestruturada, o que pode dificultar a identificação de irregularidades, conforme o trabalho de Soylu et al. (2022). Tal cenário reforça a pertinência da adoção de soluções tecnológicas, como o uso de inteligência artificial, para a sistematização e análise automatizada de contratos públicos.

2.1.4 Desafios operacionais na gestão contratual

Apesar dos avanços normativos e tecnológicos na gestão pública, ainda podem ser observados entraves operacionais relevantes. As leis federais nº 13.460/2017 (Lei de Proteção e Defesa dos Usuários de Serviços Públicos) e nº 14.129/2021 (Dispõe sobre princípios, regras e instrumentos para o Governo Digital) constituem uma tentativa relativamente recente de coibir a fragmentação das bases de dados, a ausência de padronização e a baixa interoperabilidade entre sistemas administrativos.

Ademais, o volume de contratos firmados, especialmente em áreas sensíveis como saúde e infraestrutura, aumenta a complexidade da fiscalização (OECD, 2023). Essa

circunstância pode revelar a necessidade de instrumentos capazes de auxiliar na extração e organização automática dos dados, como os modelos de linguagem de larga escala (LLMs), discutidos nas seções seguintes.

2.2 MODELOS DE LINGUAGEM DE LARGA ESCALA (LLMS) E INTELIGÊNCIA ARTIFICIAL NA GESTÃO PÚBLICA

Os Modelos de Linguagem de Larga Escala (LLMs) são inovações da IA capazes de compreender a linguagem natural para resolver tarefas complexas. O panorama da pesquisa em IA aplicada ao Setor Público abrange o potencial para uma gestão baseada em dados, mas também pode levantar questões sobre a validação da informação e a mitigação de vieses.

Os Modelos de Linguagem de Larga Escala constituem uma das principais inovações da inteligência artificial nos últimos anos. Segundo Zhao et al. (2023), os LLMs são algoritmos treinados com grandes volumes de dados textuais, sendo capazes de compreender linguagem natural e resolver tarefas complexas, como sumarização, tradução e extração de informação.

A versatilidade dos LLMs decorre de sua arquitetura baseada em “transformadores”, que permitem o aprendizado contextualizado e a geração de respostas coerentes, conforme demonstrado por VASWANI et al. (2017). Essa tecnologia vem sendo amplamente explorada na academia e na indústria, com resultados expressivos em tarefas de processamento de linguagem natural.

A utilização dos LLMs em auditoria pública tem se mostrado promissora. Li et al. (2024) demonstraram que a aplicação desses modelos em relatórios de auditoria permitiu uma acurácia de 96% na verificação de informações e reduziu em 83% o tempo de análise, quando comparado a métodos tradicionais.

A pesquisa mostra que os LLMs não apenas aceleram o processo de auditoria, como também ampliam sua profundidade, ao identificar padrões e inconsistências de forma automática. Quando aplicados à análise de contratos públicos, os modelos podem extrair dados de tabelas relacionais, identificar valores, prazos, fornecedores e cláusulas contratuais com alto grau de precisão.

O crescimento da pesquisa em inteligência artificial é evidente. O relatório AI Index 2025, publicado pela Stanford University, aponta um aumento significativo na produção científica voltada para IA, com destaque para sua aplicação em áreas de alto impacto social.

No âmbito da gestão pública, a adoção de tecnologias de IA pode auxiliar na construção de uma administração mais transparente e fundamentada em dados. A análise automatizada de contratos, nesse contexto, representa uma aplicação de como a inovação pode ser utilizada para aprimorar a eficiência governamental e a responsabilização dos agentes (OECD, 2025).

Apesar de seu potencial, o uso de LLMs apresenta desafios. Schmidt, Maiden e Maiden (2023) ressaltam que a aplicação dessa tecnologia requer cuidados quanto à validação das informações extraídas e ao combate a vieses algorítmicos. A depender do modelo utilizado e dos dados de treinamento, há riscos de omissão ou distorção de informações sensíveis.

2.3. ESTUDOS ANTERIORES

O Quadro 1 a seguir sintetiza os objetivos, métodos e resultados dos principais estudos que fundamentam esta pesquisa.

Quadro 1 – Mapeamento dos estudos anteriores relevantes

Autor (s)	**Objetivo**	**Método**	**Resultado**
Bataglia e Farranha, 2018.	O artigo tem como objetivo investigar a relação entre corrupção, controle social, transparência e acesso à informação, com foco específico na transparência passiva como instrumento para o exercício do controle social.	Pesquisa exploratória, com base empírica em dados públicos extraídos do portal da Controladoria-Geral da União (CGU), mais especificamente de pedidos de acesso à informação contendo o termo “corrupção” feitos ao Executivo Federal nos anos de 2015 e 2016.	- Nem todos os pedidos com o termo "corrupção" visavam o efetivo controle de atos corruptos, muitos apenas continham o termo incidentalmente; - Os pedidos foram direcionados tanto a órgãos da administração direta quanto indireta; - Diversas solicitações referiam-se a contratos, cláusulas anticorrupção, pedidos investigativos ou dúvidas administrativas; - Alguns pedidos eram desabafos ou manifestações políticas.

Coelho, Silva, Cunha, e Teixeira, 2018.	O artigo tem como objetivo avaliar o grau de transparência nos portais governamentais dos estados e grandes municípios brasileiros, examinando como e em que medida os entes subnacionais disponibilizam informações públicas online.	Trata-se de uma pesquisa empírica quantitativa e qualitativa, que se baseia na análise de conteúdo dos portais eletrônicos de: Todos os 26 estados brasileiros + Distrito Federal; As capitais estaduais e os municípios com população superior a 400 mil habitantes (totalizando 85 entes analisados).	- A transparência é incompleta e, em geral, focada no cumprimento mínimo das exigências legais, especialmente nas áreas de prestação de contas. - A média do IT foi de 4,58 para os estados e 4,33 para os municípios (em uma escala de 0 a 7), o que representa cerca de 65% da pontuação máxima possível, considerado um desempenho insatisfatório. - As maiores deficiências foram observadas nas categorias formas de comunicação e facilidades de uso dos portais. - Houve desigualdade regional significativa: estados e municípios do Sul e Sudeste tiveram os melhores desempenhos; os do Norte e Nordeste, os piores.
Li, Freitas, Lee e Vasarhelyi, 2024.	O artigo propõe um framework baseado em Modelos de Linguagem de Larga Escala (LLMs) para melhorar a auditoria contínua (CA) por meio da verificação cruzada em tempo real de informações contábeis com evidências textuais.	O estudo segue a metodologia de pesquisa em ciência do design (Design Science Research Methodology – DSRM) e propõe um framework em três etapas: Aquisição e pré-processamento dos dados textuais; Inferência com LLMs e verificação cruzada.	- A aplicação do framework obteve 96% de acurácia na extração de evidências auditáveis. - Houve uma redução de 83% no tempo de verificação cruzada, comparado ao processo manual tradicional. - A substituição da amostragem aleatória por testes sobre toda a população de dados aumentou significativamente o poder de controle.

			- O framework foi validado com apoio da equipe de auditoria interna (GAPES) e mostrou potencial para ser generalizado para outras áreas além da folha de pagamento, como compras públicas e contratos.
Schmidt, Maiden e Maiden, 2023.	O artigo propõe uma abordagem computacional para extrair dados fiscais contidos em relatórios de auditoria, especificamente emitidos por instituições subnacionais de controle no México, por meio de técnicas de NLP (Processamento de Linguagem Natural).	A pesquisa emprega uma metodologia aplicada de ciência de dados, com um processo de extração textual dividido em quatro etapas principais: coleta de dados, pré-processamento, classificação de texto e extração de informação.	- O pipeline reduziu de 28.000 parágrafos para 6.517 parágrafos relevantes de forma automatizada. - O modelo de classificação de texto obteve F1-score de 0,91, e o NER alcançou F1-score de 0,93. - Um conjunto de dados com 125 observações de discrepâncias fiscais extraídas de relatórios de 18 municípios, cobrindo os anos de 2008 a 2016. - As discrepâncias totalizaram, em média, 4% do orçamento municipal, chegando a 26% em alguns casos.
Vaswani et al., 2017.	O artigo propõe uma nova arquitetura de modelo de transdução de sequência, chamada Transformer, que elimina totalmente o uso de redes neurais recorrentes (RNNs) e convolucionais, substituindo-as por mecanismos de atenção (attention), especificamente self-attention.	Teste do modelo que segue o padrão encoder-decoder, onde: O encoder processa toda a entrada com atenção própria (self-attention); O decoder utiliza atenção tanto sobre a entrada (encoder-decoder attention) quanto sobre sua própria saída passada	- Na tarefa WMT 2014 English-to-German, o Transformer (modelo “big”) alcançou 28.4 BLEU, superando os modelos anteriores, inclusive ensembles. - Para English-to-French, atingiu 41.0 BLEU, com tempo de treinamento significativamente menor (3.5 dias em 8 GPUs), a uma fração do

		(masked self-attention), garantindo a propriedade autoregressiva.	custo computacional dos modelos anteriores.
Zhao et al., 2023.	O artigo tem como objetivo oferecer uma revisão sistemática e abrangente sobre os Modelos de Linguagem de Larga Escala (LLMs).	O artigo utiliza o método de revisão de literatura técnica, sintetizando os principais estudos, benchmarks, frameworks e práticas de engenharia relacionadas aos LLMs.	- LLMs (como GPT-4) desenvolvem habilidades complexas que não aparecem em modelos menores, como raciocínio passo a passo, seguimento de instruções e aprendizado contextualizado (*in-context learning*). - O desempenho dos modelos cresce significativamente com mais dados, parâmetros e poder computacional, seguindo padrões como as leis de escala de Kaplan e Chinchilla. - LLMs modernos são aplicáveis a uma ampla gama de tarefas – texto, código, multimodalidade – com alta eficácia, superando benchmarks anteriores.

Fonte: elaborado pelo autor.

Os estudos iniciais (Bataglia & Farranha, 2018; Coelho et al., 2018) tiveram como foco a identificação e a mensuração das deficiências na governança pública brasileira. Os autores apontam que a transparência é, de modo geral, reativa e incompleta; os portais se limitam a cumprir o mínimo legal, o que dificulta o acesso e a utilização da informação pelo cidadão. O acesso à informação, embora concebido como instrumento de controle social, mostra-se frequentemente ineficaz para a fiscalização de atos corruptos, revelando uma lacuna entre a norma legal e sua aplicação prática.

Paralelamente, o quadro apresenta os pilares tecnológicos que permitem superar esses desafios. O artigo de Vaswani et al. (2017) é fundamental, pois introduz a arquitetura Transformer, a inovação que possibilitou o surgimento dos modelos de linguagem modernos. Anos depois, o trabalho de Zhao et al. (2023) consolida esse avanço, pois oferece uma revisão

completa sobre os LLMs e confirma suas capacidades avançadas de raciocínio, aprendizado contextual e aplicação em tarefas complexas, o que os torna ferramentas maduras e prontas para uso.

Os artigos mais recentes (Schmidt et al., 2023; Li et al., 2024) contribuem significativamente ao aplicar diretamente as tecnologias de IA (NLP e LLMs) aos problemas de controle e auditoria. Eles demonstram na prática como a IA pode automatizar a extração e verificação de dados em larga escala, tarefas que seriam manuais, lentas e baseadas em amostragem. Os resultados são expressivos: redução drástica no tempo de verificação (83% no caso de Li et al.), alta acurácia (91-96%) e a capacidade de analisar 100% dos dados, o que aumenta de forma significativa o poder de controle e fiscalização.

3 METODOLOGIA

3.1. CLASSIFICAÇÃO DA PESQUISA

Este estudo é caracterizado como uma pesquisa descritiva, com uma abordagem quali-quantitativa. Sua natureza descritiva, conforme a definição de Gil (2022), manifesta-se na finalidade de observar, registrar e analisar as características de um fenômeno particular sem manipulá-lo. Neste trabalho, a pesquisa realiza uma varredura integral dos contratos públicos em PDF, contemplando conteúdos estruturados e não estruturados (texto corrido, tabelas, quadros, campos rotulados e metadados legíveis), a fim de descrever a verificabilidade da estrutura e integridade documental e avaliar a performance de um modelo de linguagem de larga escala na extração automatizada das informações. O desempenho é mensurado por acurácia (precisão da extração) e eficiência (tempo médio de extração por contrato).

De acordo com Creswell e Creswell (2022), a abordagem mista (quali-quantitativa) é adotada para proporcionar uma compreensão mais completa do problema. No quantitativo, adota-se a mensuração objetiva do desempenho do extrator por duas métricas: acurácia (entendida como a precisão da extração dos campos definidos) e eficiência (o tempo médio de extração por contrato). No qualitativo, procede-se à análise de conteúdo para interpretar a estrutura e integridade dos contratos, bem como a coerência e completude dos dados extraídos.

3.2 COLETA DE DADOS

Para atender à natureza descritiva e à abordagem mista do estudo, o procedimento de coleta de dados foi estruturado em duas frentes. A primeira define o universo documental a ser investigado, estabelecendo a população e a amostra da pesquisa. A segunda detalha as ferramentas tecnológicas que foram empregadas para a obtenção e extração das informações desses documentos (Gil, 2008).

3.2.1 População e amostra

A população deste estudo compreende o universo de contratos públicos firmados pelo Departamento de Logística em Saúde (DLOG) e disponibilizados publicamente no portal do Ministério da Saúde do Brasil, acessível no endereço eletrônico: https://www.gov.br/saude/pt-br/acesso-a-informacao/licitacoes-e-contratos/contratos-dlog.

A amostra foi definida de forma não probabilística, por conveniência, sendo composta pelos 100 primeiros contratos disponibilizados para acesso e download no referido portal, referentes ao ano de 2024. Foram incluídos na análise todos os documentos selecionados, e a extração automatizada contemplou: Número do Contrato, Data da Assinatura, Objeto, Fornecedor, CNPJ, Valor Total, Moeda, Prazo de Vigência (meses) e Código de Elemento de Despesa.

3.2.2 Arquitetura tecnológica para extração de dados

A extração de informações dos documentos foi realizada por meio de uma arquitetura tecnológica que integra um Modelo de Linguagem de Larga Escala (LLM) e uma plataforma de planilhas em nuvem do Google. O ecossistema definido para este trabalho é composto por:

Quadro 2 – Definição do modelo de linguagem e do ambiente de armazenamento

Modelo de Linguagem	**Ambiente de Execução e Armazenamento**
Foi utilizado o Gemini 2.5 Pro, desenvolvido pelo Google. Este modelo foi escolhido por sua capacidade multimodal de processar grandes volumes de documentos, incluindo arquivos em formato PDF, e sua habilidade em seguir instruções complexas para extração e estruturação de dados, no entanto, a opção “deep research” não estará ativada para evitar pesquisa em fontes externas.	Os dados extraídos pelo modelo foram organizados e armazenados no Google Planilhas. Esta ferramenta foi selecionada por sua acessibilidade, facilidade de uso para análise de dados tabulares e capacidades de colaboração.

Fonte: elaborado pelo autor.

Para a execução da extração, cada um dos 100 contratos em formato PDF foi submetido ao modelo Gemini 2.5 Pro de forma individualizada. O procedimento consistiu em

abrir um chat de conversação dedicado para cada documento, no qual o arquivo PDF era anexado seguido imediatamente pela inserção do prompt estruturado. O modelo processou o documento e retornou os dados no formato JSON exigido, e esse processo foi repetido sequencialmente, contrato por contrato. A saída do modelo, em formato de dados estruturados, foi então transposta para o Google Planilhas, onde cada linha passou a representar contrato e cada coluna, uma das variáveis de interesse, viabilizando a análise quantitativa descrita na seção 3.3.

3.2.3 O Instrumento de extração: o prompt estruturado

O principal instrumento para a coleta de dados é um prompt (um conjunto de instruções cuidadosamente elaborado) que orienta o LLM na execução da tarefa. O desenvolvimento de um prompt eficaz é fundamental para garantir a acurácia e a consistência da extração (Xu et al., 2024). O prompt utilizado nesta pesquisa foi projetado com uma estrutura multicomponente, conforme o Guia de Engenharia Prompt da DAIR.AI, em que cada parte desempenha uma função específica para guiar o comportamento do modelo (DAIR.AI, 2025). As seções a seguir detalham cada um desses componentes.

O prompt foi segmentado em blocos lógicos para assegurar que o modelo compreenda o contexto, a tarefa, o formato de saída e as regras de conduta.

3.2.3.1 Definição de papel

O texto inicia com a atribuição de um papel específico ao LLM:

"Assuma o papel de um Modelo de Linguagem de Larga Escala (LLM) especializado em extração, organização e verificação de informações em contratos públicos."

Esta instrução contextualiza o modelo, fazendo com que ele ative o conhecimento relevante para o domínio de documentos legais e administrativos, o que tende a aumentar a precisão na identificação dos dados solicitados.

3.2.3.2 Comando de extração e definição das variáveis

Em seguida, o comando principal é emitido, listando explicitamente os campos (variáveis) a serem extraídos do documento:

"Extraia os seguintes dados do contrato público anexado [n° do contrato, data da assinatura, objeto (apenas o medicamento ou equipamento/serviço), fornecedor, CNPJ, valores, moeda, prazo de vigência e código de elemento de despesa]"

O objetivo desta instrução é delimitar o escopo da tarefa e explicitar os campos a serem extraídos, buscando reduzir ambiguidades semânticas e variações de interpretação. O direcionamento do modelo para rótulos e padrões específicos de contratos administrativos também favorece a completude das respostas. Adicionalmente, esta abordagem contribui para a padronização do procedimento e para a futura validação cruzada dos resultados.

3.2.3.3 Estruturação da saída de dados (Esquema em JSON)

Para garantir que a saída seja consistente e apta a ser importada e tratada em planilhas eletrônicas, o prompt define um esquema em formato JSON (JavaScript Object Notation), um padrão leve de troca de dados, de fácil leitura e escrita tanto para humanos quanto para máquinas:

“{

"contratos": [

{

"numero": "string",

"data_assinatura": "DD/MM/AAAA",

"objeto": "string",

"fornecedor": "string",

"cnpj": "00.000.000/0000-00",

"valores": {

"valor_total": 0.0,

"moeda": "BRL"

},

"prazo_vigencia_meses": "00",

"codigo_elemento_despesa": "string"

}

]

}”

A utilização de um esquema pré-definido força o modelo a entregar os dados em uma estrutura previsível e padronizada, o que facilita a importação para o Google Planilhas e a posterior análise.

3.2.3.4 Regras de formatação e comportamento

Por fim, são adicionadas regras para lidar com casos específicos e para restringir o comportamento do modelo, isto com o intuito de dirimir erros comuns como a invenção de dados:

Formatação de texto longo:

*"Se qualquer dado contiver um texto muito longo (como o nome do fornecedor), insira quebras de linha manual com a tag de html
”.*

Esta regra visa a fazer com que os resultados da extração não ultrapassem suas colunas correspondentes.

Tratamento de Inconsistências:

"Leve em conta inconsistências, riscos e lacunas, mas traga os dados de forma idônea.".

Esta instrução orienta o modelo a ser fiel ao documento-fonte, mesmo que as informações nele contidas pareçam incompletas ou ambíguas.

Restrição contra "alucinações":

"Não invente dados inexistentes.".

Esta é uma diretriz fundamental para assegurar a integridade da pesquisa, pois ela instrui o modelo a deixar um campo em branco caso a informação não seja encontrada no documento.

Condições de contexto e finalização:

"Considere que o contrato está em PDF";

"Finalize quando o documento tiver sido processado".

Essas condições definem o ambiente e o ponto de parada da extração. Ao fixar que a entrada é um PDF, o prompt estabelece pressupostos técnicos para reduzir ambiguidades na leitura. A instrução para finalizar quando o documento for processado cria um critério de término objetivo e evita margem para sugestões indesejadas.

3.2.4 Procedimento de verificação de OCR (dados qualitativos coletados)

Com o objetivo de interpretar condições documentais que podem afetar a extração automatizada, cada contrato em PDF foi avaliado quanto à presença e qualidade de texto extraível (OCR). Para cada documento, foram coletados os dados dos elementos indicados no item 3.3.

Os elementos foram produzidos automaticamente a partir do texto extraído do PDF através de um código executado no Google Colab em linguagem Python por bibliotecas voltadas a análise de PDFs (primeiro por PyMuPDF e, em *fallback*, pdfplumber), sem editar o conteúdo do arquivo. O código está disponível no Apêndice A deste trabalho.

Visando uma melhor organização do processo, os 100 contratos foram armazenados em uma pasta única e processados em ordem alfabética natural. Os nomes originais dos PDFs foram preservados em coluna própria (“arquivo_pdf”) da base.

3.3 TRATAMENTO DE DADOS E ANÁLISE DE DADOS

A análise dos dados coletados foi realizada por meio de uma abordagem mista, que combina técnicas quantitativas e qualitativas para proporcionar uma compreensão do fenômeno estudado (Creswell e Clark, 2017). Este método permitiu mensurar o desempenho da extração de dados, e interpretar as nuances e os desafios práticos envolvidos. A capacidade do modelo de IA de interpretar e organizar dados complexos de documentos administrativos foi o foco central desta análise. O tratamento dos dados foi dividido conforme as duas vertentes da pesquisa:

Análise quantitativa: Esta etapa concentrou-se na avaliação numérica da efetividade do modelo de inteligência artificial. O modelo foi programado para extrair automaticamente os dados previamente definidos de cada contrato individualmente. A mensuração do desempenho ocorreu por meio de duas variáveis principais: a acurácia e a eficiência.

A acurácia correspondeu ao percentual de sucesso na extração dos campos. Para aferir esta métrica, foi realizada uma verificação manual de todos os dados extraídos, comparando-os, campo a campo, com um gabarito criado a partir da leitura de cada contrato. Para o campo "Data da Assinatura", foi considerado um acerto caso o modelo extraísse qualquer uma das datas presentes, visto que os signatários podem assinar em dias distintos e o prompt não especifica qual data extrair. Quanto ao "Objeto" do contrato, como não há um nível de detalhamento predefinido para a extração, a acurácia foi medida por um percentual de acerto, calculado pela fórmula =LEN(célula_extraída) / LEN(célula_manual) x 100, que compara a quantidade de caracteres extraídos pelo modelo com a quantidade de caracteres do objeto transcrito manualmente no gabarito. A eficiência foi avaliada pelo tempo médio de processamento dos contratos, tendo o período cronometrado do momento em que se comanda o LLM a iniciar a tarefa até o momento em que a tarefa é finalizada. Adicionalmente, para estabelecer um padrão de comparabilidade e dimensionar o ganho de eficiência, foi realizada no mesmo ambiente computacional da extração automática uma extração manual de dados em cinco contratos da amostra que foram escolhidos aleatoriamente. O tempo médio gasto por um

profissional para extrair as nove variáveis predefinidas nesses documentos foi cronometrado em 1 minuto e 53 segundos (113 segundos) desde o download do arquivo à conclusão da tarefa.

Os resultados obtidos foram dispostos através de estatística descritiva, utilizando medidas de tendência central (média e mediana), dispersão (desvio-padrão e amplitude) e distribuição de frequência. Essa abordagem permitiu comparar o desempenho da tecnologia em possíveis diferentes cenários e avaliar seu potencial de aplicação na tarefa de extração automatizada de dados em contratos públicos, sem pressupor conclusões prévias.

Análise Qualitativa: Esta etapa focalizou a identificação e a qualidade do OCR presente nos contratos em PDF, por compreender que a disponibilidade e a integridade do texto extraível impactam diretamente a extração automática. Cada documento foi auditado conforme indicado na tabela abaixo:

Quadro 3 – Detalhamento dos indicadores de observabilidade textual

Elemento analisado	**Significado**	**Como o código calcula**
COP_% (Cobertura OCR por página)	Percentual de páginas com texto selecionável no PDF.	COP_% = (páginas com texto / páginas totais) x 100
PRC_% (Precisão de Rótulos Canônicos)	Percentual de rótulos-chave localizáveis via busca nativa (“CNPJ”, “Objeto”, “Valor Total”, “Vigência”, “Elemento de Despesa”).	PRC_% = (soma de rótulos encontrados / número de rótulos buscados) x 100
CFN_% (Coerência de Formatação Numérica)	Proporção de amostras textuais em que números e valores monetários preservam separadores e decimais de forma consistente.	Escolhe 3 trechos, de início, meio e fim do documento, limitado a 1500 caracteres cada. Avalia o trecho como coerente se, existem grupos de dígitos plausíveis: \b\d{1,3}(?:\.\d{3})*,\d{2}\b ou \d{3,}[\.,]?\d*

		CFN_% = (número de trechos coerentes / total de trechos analisados) x 100
RC_% (Ruído de Caracteres)	Percentual de glifos problemáticos (exemplo: "  ", caracteres de controle) nas amostras de texto.	Utiliza as mesmas 3 amostras do CFN_% e procura caracteres problemáticos ou incomuns. RC_% = (caracteres ruins / caracteres totais) x 100
ITB_0ou1 (Integridade de Tabelas – heurística)	Indicação binária de presença de estrutura de tabela reconhecível no texto.	Busca pela tabulação "\t", ou múltiplos blocos de espaço, ou muitos delimitadores ";"/"," na mesma linha e soma pontos. Se a pontuação for ≥ 5 linhas ou ≥ 2% das linhas, o que for maior, ITB = 1; senão, 0.
BVK_0a1 (Busca de Valores-Chave)	Presença (1), presença parcial (0,5) ou ausência (0) de CNPJ e valor monetário detectados por expressões regulares.	Busca por regex (expressões regulares) por: • CNPJ, sendo \b\d{2}\.?\d{3}\.?\d{3}/?\d{4}-?\d{2}\b • Moeda, sendo \b\d{1,3}(?:\.\d{3})*,\d{2}\b Se encontra ambos, atribui 1, se encontra apenas um, atribui 0,5, se nenhum, 0.
OCRScore (0–100)	Índice sintético que compila os indicadores anteriores para resumir a utilizabilidade do OCR no documento. O peso atribuído a cada variável no cálculo é	OCRScore = 0,20 x COP + 0,25 x PRC + 0,20 x CFN + 0,10 x (100 – RC) + 0,15 x (ITB x 100) + 0,10 x (BVK x 100)

	deliberado com base no papel funcional que elas exercem.	
Nível_OCR (N0–N3)	Classificação derivada do OCRScore (N0: crítico; N1: fraco; N2: funcional; N3: ótimo).	0 a 24 → **N0** (crítico) 25 a 49 → **N1** (fraco) 50 a 79 → **N2** (funcional) 80 a 100 → **N3** (ótimo)

Fonte: elaborado pelo autor.

O procedimento de análise da qualidade documental foi executado de forma automatizada, utilizando um código em linguagem Python para identificar e mensurar as características textuais de cada contrato em PDF. Este processo calculou, para cada documento, os indicadores definidos na metodologia, como a Precisão de Rótulos Canônicos (PRC_%), a Coerência de Formatação Numérica (CFN_%) e o Ruído de Caracteres (RC_%).

Os resultados desta análise foram consolidados e apresentados no Capítulo 4 por meio de estatística descritiva (médias, medianas e desvios-padrão dos indicadores) e pela distribuição de frequência dos contratos segundo o Nível de OCR (N0-N3).

Essa avaliação da qualidade documental foi desenhada para complementar a análise quantitativa de desempenho. O objetivo é fornecer um contexto técnico para interpretar as variações de acurácia e eficiência, explicitando condições documentais que poderiam favorecer ou dificultar a extração automatizada (Hamdi et al., 2023). Conforme discutido nos resultados, essa análise foi crucial para revelar a resiliência do modelo de linguagem, cuja performance se mostrou largamente independente das métricas de qualidade textual observadas.

4 APRESENTAÇÃO E DISCUSSÃO DOS RESULTADOS

Neste capítulo, são apresentados e interpretados os resultados da extração de dados aplicada à amostra de contratos, organizados em três eixos de análise: o desempenho quantitativo do modelo (acurácia e eficiência), a qualidade documental dos arquivos (observabilidade textual) e a relação entre essas duas dimensões.

A análise inicia-se com a avaliação do desempenho de extração, detalhando a acurácia por campo, que se mostrou praticamente perfeita na maioria das variáveis, e a alta eficiência do processo, medida pelo tempo de processamento por contrato. Em seguida, o capítulo dedica-se ao diagnóstico da qualidade documental, caracterizando a amostra por meio de um conjunto de métricas de OCR e classificando os contratos em níveis de observabilidade textual.

Dado que a análise estatística é responsável por conectar os dois eixos anteriores (performance e qualidade), o ponto central reside na investigação, por meio de correlação estatística, da associação entre a qualidade documental e o desempenho do modelo, o que revela uma notável independência do LLM em relação às imperfeições textuais dos arquivos.

4.1 CARACTERÍSTICAS ACIDENTAIS RELEVANTES

As propriedades físicas e digitais dos PDFs são aqui denominadas "características acidentais". Elas incluem, por exemplo, a origem do documento (nativo/digitalizado), resolução, compressão, orientação, a camada de texto, codificação de caracteres, presença de ruído gráfico, formatação numérica e a integridade de estruturas tabulares. Tais características podem afetar a observabilidade textual e, consequentemente, os resultados de acurácia e eficiência da extração (FADGI, 2023). Esta subseção, portanto, descreve as propriedades observadas na amostra e apresenta os indicadores definidos para verificar esses aspectos.

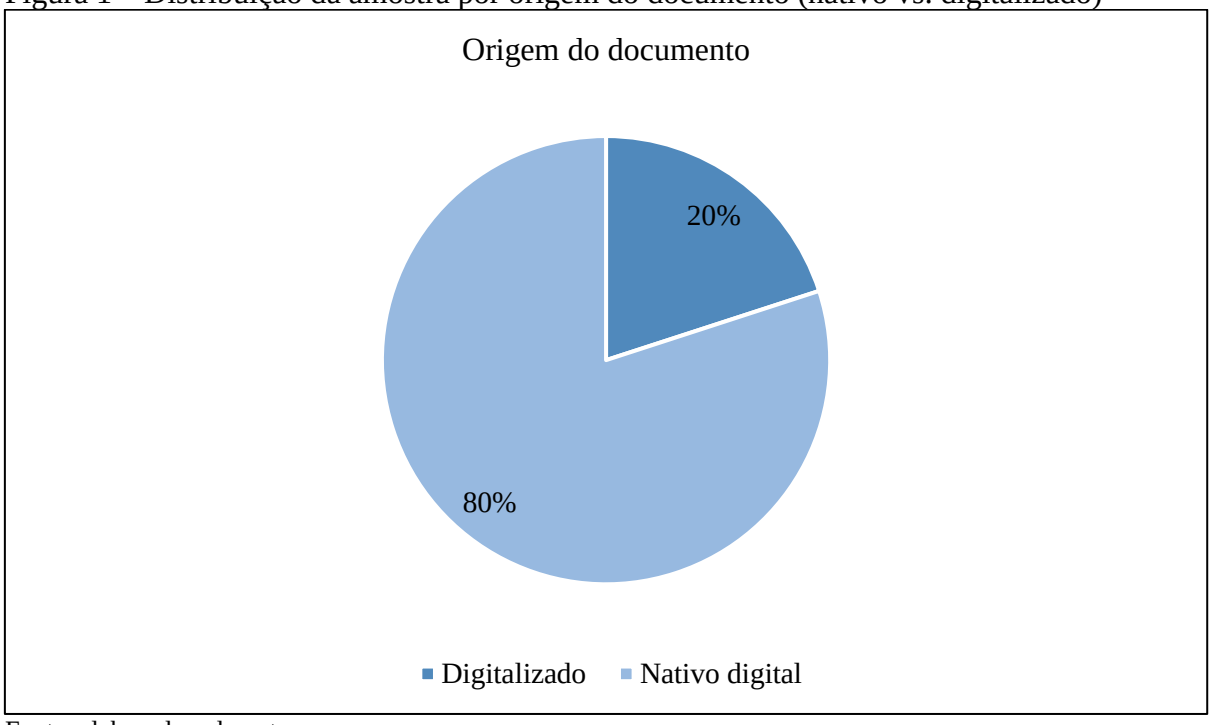
4.1.1 Origem dos documentos

Neste trabalho, entende-se por “documento nativo digital” o PDF gerado por software, contendo camada de texto selecionável; “documento digitalizado” refere-se ao PDF composto por imagens, cujo texto é tornado pesquisável via OCR.

Para fins de contextualização da amostra, é relevante notar que 80% dos contratos analisados são nativos digitais e 20% são digitalizados. Esta distribuição ilustra a diversidade dos arquivos encontrados em portais públicos. Contudo, para os fins deste estudo, a origem do documento não foi tratada como uma variável nas análises de desempenho subsequentes, servindo apenas para caracterizar o conjunto de documentos da pesquisa.

A Figura 1 ilustra visualmente esta distribuição percentual da amostra.

Figura 1 – Distribuição da amostra por origem do documento (nativo vs. digitalizado)



Fonte: elaborado pelo autor.

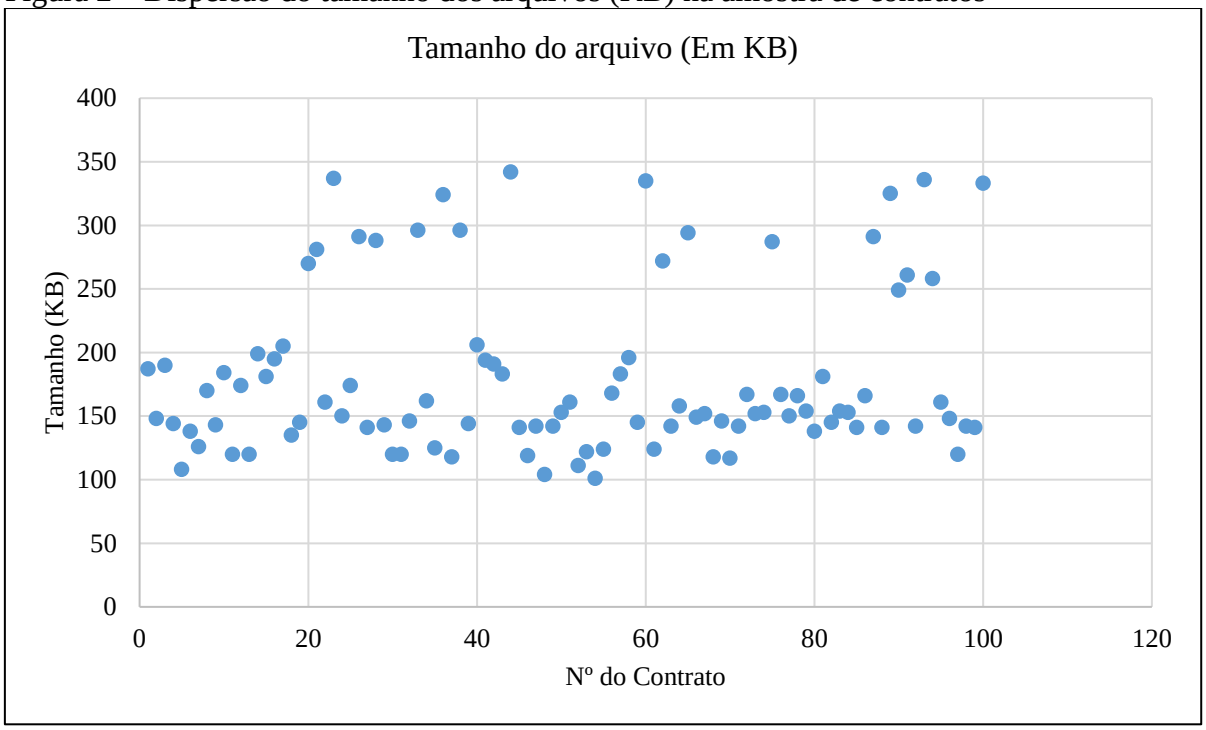
4.1.2 Tamanho dos arquivos (em KB)

Outra característica acidental da amostra é o tamanho dos arquivos em kilobytes (KB). Na amostra analisada, o tamanho variou entre 101 KB e 342 KB. Embora o tamanho do arquivo possa, em teoria, estar associado a fatores como o número de páginas, a resolução de imagens ou o nível de compressão, esta variável é apresentada aqui com o propósito exclusivo de contextualização.

Ela ajuda a dimensionar as propriedades dos documentos digitais que compõem o universo estudado. Deste modo, assim como a origem do documento, o tamanho do arquivo não foi empregado como variável nas análises de correlação que se seguem, cumprindo um papel meramente descritivo.

A dispersão do tamanho dos arquivos (KB) na amostra de contratos é apresentada na Figura 2.

Figura 2 – Dispersão do tamanho dos arquivos (KB) na amostra de contratos



Fonte: elaborado pelo autor.

4.2 DESEMPENHO QUANTITATIVO DE EXTRAÇÃO (ACURÁCIA E EFICIÊNCIA)

Esta seção apresenta os resultados quantitativos centrais da pesquisa, focando nas duas principais métricas de desempenho estabelecidas na metodologia: a acurácia, que mensura "o quanto" o modelo acerta na extração dos dados, e a eficiência, que avalia "quão rápido" a tarefa é executada.

4.2.1 Acurácia

O desempenho do modelo LLM na extração dos campos definidos demonstrou um nível de precisão e confiabilidade extremamente elevado. A análise revelou que oito dos nove campos extraídos alcançaram uma acurácia perfeita de 100,00% em toda a amostra de 100 contratos. Os campos com tal performance foram: “Nº do Contrato”, “Data da Assinatura”, “Fornecedor”, “CNPJ”, “Valor Total”, “Moeda”, “Prazo de Vigência (meses)” e “Código de Elemento de Despesa”. Este resultado evidencia a capacidade do modelo de identificar e extrair dados estruturados ou semiestruturados com exatidão absoluta, independentemente das variações documentais encontradas.

O campo “Objeto” foi a única exceção ao desempenho de 100%. Um exemplo da não consecução de extração perfeita é o que ocorreu com o contrato 57/2024, em que o LLM retornou o objeto “COMPRESSA GAZE” em vez de “COMPRESSA GAZE, TECIDO 100% ALGODÃO, 13 FIOS/CM2, COR BRANCA, ISENTA DE IMPUREZAS, 8 CAMADAS, 7,50 CM, 7,50 CM, 5 DOBRAS, DESCARTÁVEL” acertando apenas 14 dos 133 caracteres, ou seja, 10,53%.

Por se tratar de um campo textual de formato livre e extensão variável, sua extração apresentou um comportamento distinto, atingindo uma acurácia média de 91,13%. A performance neste campo variou, com um valor mínimo registrado de 8,85% e um máximo de 100,00%. Tal variabilidade se explica pela própria natureza do dado e pela metodologia de cálculo utilizada, que comparou a quantidade de caracteres da extração automatizada com a transcrição manual. Por esse critério, extrações que capturavam a essência do objeto, mas omitiam detalhes secundários, foram penalizadas, resultando em valores de acurácia inferiores a 100%.

A Tabela 1 apresenta a estatística descritiva detalhada da acurácia de extração para cada um dos nove campos analisados.

Tabela 1 – Estatística descritiva da acurácia de extração do LLM por campo

Campo	Média	Mediana	Desvio-padrão P	Amplitude
Número do Contrato	100,00%	100,00%	0	0,00%
Data da Assinatura	100,00%	100,00%	0	0,00%

Objeto	91,13%	100,00%	0,2470	91,15%
Fornecedor	100,00%	100,00%	0	0,00%
CNPJ	100,00%	100,00%	0	0,00%
Valor Total	100,00%	100,00%	0	0,00%
Moeda	100,00%	100,00%	0	0,00%
Prazo de Vigência (meses)	100,00%	100,00%	0	0,00%
Código de Elemento de Despesa	100,00%	100,00%	0	0,00%

Fonte: elaborado pelo autor.

Ainda assim, uma média superior a 90% para o campo mais complexo e não estruturado do conjunto, somada à precisão absoluta nos demais campos, atesta a alta efetividade e o grau de êxito do modelo na tarefa designada.

4.2.2 Eficiência

A eficiência do processo foi avaliada pelo tempo médio de processamento por contrato, cronometrado em segundos. Os resultados demonstram um desempenho notavelmente rápido e consistente.

O tempo médio de extração para um único contrato foi de 19,53 segundos, com uma mediana muito próxima de 19,75 segundos. A consistência do processo é evidenciada pelo baixo desvio padrão de 2,42 segundos, indicando que o tempo de execução variou pouco entre os diferentes documentos. O contrato mais rápido foi processado em 14,95 segundos e o mais lento em 23,92 segundos.

A performance de extração do LLM pode ser comparada ao tempo de 113 segundos gastos na extração manual de dados de um contrato, conforme estabelecido na metodologia. Em termos relativos, o tempo médio de 19,54 segundos representa um ganho de eficiência de 82,7% em relação ao processo manual tradicional. Tal resultado evidencia que o LLM acelera a tarefa de extração de dados neste contexto em mais de cinco vezes ($113 / 19{,}54 \cong 5{,}78$).

Os dados estatísticos completos referentes à eficiência (tempo de processamento) são consolidados na Tabela 2.

Tabela 2 – Estatística descritiva da eficiência (tempo de processamento) por contrato

Campo	Média	Mediana	Desvio-padrão P	Amplitude	Mínimo	Máximo
Tempo de processamento (segundos)	19,54	19,68	2,12	8,97	14,95	23,92

Fonte: elaborado pelo autor.

Cumpre ressaltar, conforme discutido na seção 4.4, que esta performance de alta eficiência não foi significativamente degradada por documentos de menor qualidade textual, o que permite inferir a escalabilidade da tecnologia para aplicação em grandes volumes de arquivos públicos.

4.3 QUALIDADE DOCUMENTAL (OCR) E OBSERVABILIDADE TEXTUAL

A efetividade de qualquer processo de extração automatizada de dados depende fundamentalmente da qualidade do material-fonte. Nesta seção, a análise volta-se para as características intrínsecas dos documentos, mensurando a qualidade do Reconhecimento Óptico de Caracteres (OCR) e a observabilidade textual geral da amostra. O intuito é diagnosticar a "saúde" documental dos contratos, identificando o grau de adequação dos arquivos para a análise computacional antes mesmo da intervenção do LLM. Para tal, utilizam-se os indicadores definidos na metodologia, cujos resultados descritivos são apresentados a seguir.

O principal indicador sintético, o OCRScore, que compila as diversas métricas de qualidade em uma pontuação de 0 a 100, apresentou uma média de 88,40 para a amostra, com um desvio padrão de 14,88. Este valor médio elevado sugere, em uma primeira análise, uma boa qualidade geral dos documentos. A pontuação mínima registrada foi de 43,80, enquanto a máxima atingiu 99,90, indicando que, embora a maioria dos arquivos seja de alta qualidade, ainda existe uma variabilidade considerável.

A distribuição dos contratos segundo o Nível_OCR (N0-N3) confirma essa tendência. Conforme detalhado na Tabela 3, a vasta maioria dos documentos (80,00%) foi classificada no nível N3 (Ótimo), caracterizados por texto claro, pesquisável e bem estruturado. Em contrapartida, uma minoria de 5,00% dos contratos enquadrou-se no nível N1 (Fraco),

representando os maiores desafios à extração automatizada, não sendo identificado nenhum contrato no nível N0 (Crítico).

A Tabela 3 resume os resultados da estatística descritiva para os principais indicadores de qualidade documental (OCR) da amostra.

Tabela 3 – Estatística descritiva dos indicadores de qualidade documental (OCR) da amostra

Elemento	Média	Mediana	Desvio-padrão P	Amplitude	Mínimo	Máximo
COP_%	100,00	100,00	0,00	0,00	100,00	100,00
PRC_%	79,80	100,00	40,15	100,00	0,00	100,00
CFN_%	94,00	100,00	18,58	100,00	0,00	100,00
RC_%	9,75	1,90	19,98	77,70	1,20	78,90
ITB_0ou1	1,00	1,00	0,00	0,00	1,00	1,00
BVK_0a1	0,84	1,00	0,34	1,00	0,00	1,00
OCRScore (0-100)	91,17	99,80	17,24	62,40	37,50	99,90

Fonte: elaborado pelo autor.

Complementarmente, a Tabela 4 detalha a distribuição de frequência absoluta dos contratos com base nessa classificação de níveis (N0-N3).

Tabela 4 – Distribuição de frequência dos contratos por nível de OCR

Nivel_OCR_(N0-N3)	
N0	0
N1	5
N2	15
N3	80

Fonte: elaborado pelo autor.

Analisando os indicadores individuais, a Precisão de Rótulos Canônicos (PRC_%) obteve uma média de 79,80%, indicando que, na maioria dos contratos, termos-chave como "CNPJ", "Objeto" e "Valor Total" são localizáveis através de busca textual simples. A Coerência de Formatação Numérica (CFN_%) registrou média de 94,00%, mostrando um alto nível de consistência na apresentação de valores. Por fim, o indicador Inverso do Ruído de Caracteres (RC_inv) foi utilizado para medir a limpeza do texto. Esta métrica representa a inversão do percentual de ruído original (100% - RC_%), para alinhar-se à lógica dos demais indicadores, onde um valor maior representa uma melhor qualidade. O RC_inv alcançou uma

média elevada de 90,25%, sinalizando que a ocorrência de glifos problemáticos ou caracteres de controle é relativamente baixa no conjunto da amostra.

Em suma, o diagnóstico da qualidade documental revela um cenário predominantemente positivo, com 80% dos contratos públicos analisados apresentando um padrão ótimo de observabilidade textual. No entanto, a existência de um subconjunto de documentos com qualidade inferior (nível N1) justifica a necessidade de ferramentas como os LLMs, capazes de lidar com as imperfeições e a heterogeneidade inerentes aos arquivos do setor público.

4.4 RELAÇÃO ENTRE DESEMPENHO E QUALIDADE DOCUMENTAL

Para investigar a associação entre as características acidentais dos documentos e o desempenho do LLM, procedeu-se ao cálculo do coeficiente de correlação de Spearman. Este método foi escolhido por sua conveniência em avaliar a força e a direção de relações monotônicas, sendo adequado para os dados em questão. É importante notar que, diferentemente da correlação de Pearson que utiliza os valores brutos, o coeficiente de Spearman opera sobre os postos (ranks) das variáveis. Ou seja, os dados são primeiro ordenados e ranqueados em relação aos demais valores da amostra, e a correlação é então calculada sobre essas posições, o que torna o método resistente a outliers e a distribuições não normais. Por essa razão, os gráficos de dispersão apresentados (Figuras 3-14) servem primariamente para uma inspeção visual da relação, enquanto o cálculo do coeficiente em si foi realizado sobre os postos dos dados, e não sobre os valores primários obtidos. Para complementar a análise dos coeficientes, foram gerados 12 gráficos de dispersão, um para cada par de variáveis (indicadores de qualidade vs. acurácia e indicadores de qualidade vs. tempo), permitindo a inspeção visual de possíveis padrões e tendências

Antes de apresentar os resultados, cumpre notar que duas variáveis da análise qualitativa, COP_% (Cobertura OCR por Página) e ITB_0oul (Integridade de Tabelas), foram omitidas nesta etapa de correlação. A omissão justifica-se pelo fato de que ambas apresentaram variância nula na amostra estudada, todos os 100 contratos registraram 100% de cobertura de OCR e indicaram a presença de estrutura tabular (valor 1), resultando em um desvio padrão de zero. Uma vez que a correlação mede a covariância entre variáveis, a ausência de variação em COP_% e ITB_0oul as torna estatisticamente inertes para esta análise.

Dessa forma, a análise de correlação concentrou-se nas variáveis que apresentaram variabilidade. Os resultados, sintetizados na Tabela 5, revelam as correlações entre os principais indicadores de qualidade documental e as duas métricas de desempenho (acurácia global e tempo de extração).

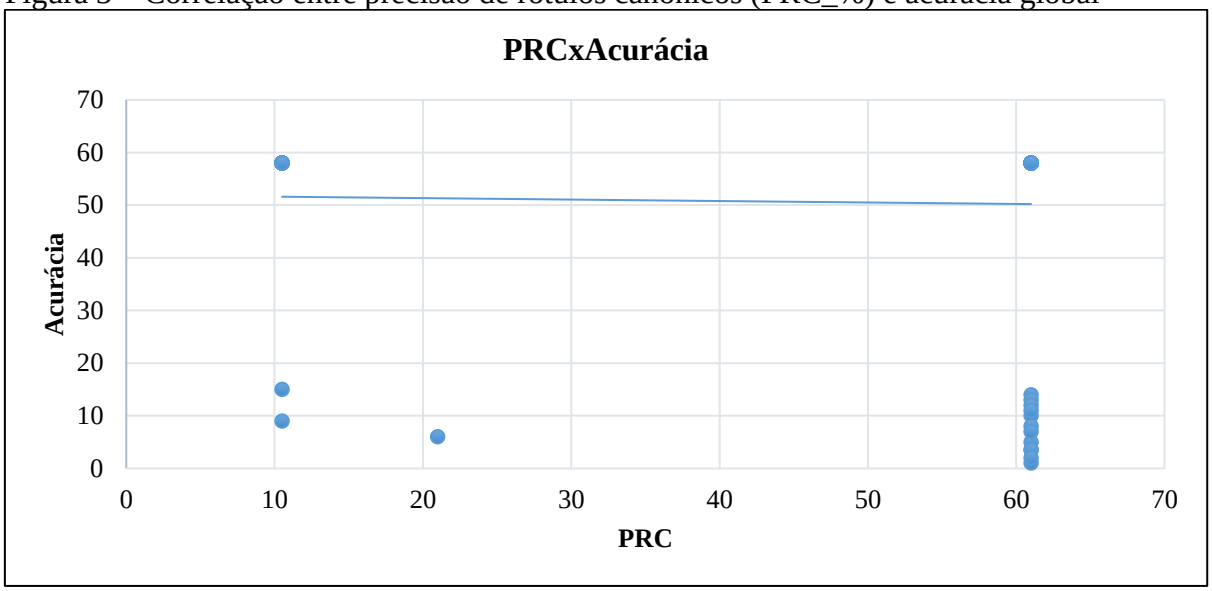
Tabela 5 – Coeficientes de correlação (Spearman) entre qualidade documental e desempenho da extração

Indicador de Qualidade	**Correlação com Acurácia Global (%)**	**Correlação com Tempo (s)**
PRC_% (Precisão de Rótulos Canônicos)	-0,031	-0,018
CFN_% (Coerência de Formatação Numérica)	-0,147	-0,011
RC_inv (Inverso do Ruído de Caracteres)	-0,134	-0,095
BVK_0a1 (Busca de Valores-Chave)	-0,084	-0,049
OCRScore_0a100 (Índice Sintético)	-0,094	-0,088
Nível_OCR (N0 a N3)	-0,087	-0,019

Fonte: elaborado pelo autor.

A Figura 3 apresenta o gráfico de dispersão que correlaciona a Precisão de Rótulos Canônicos (PRC_%) com a acurácia global da extração.

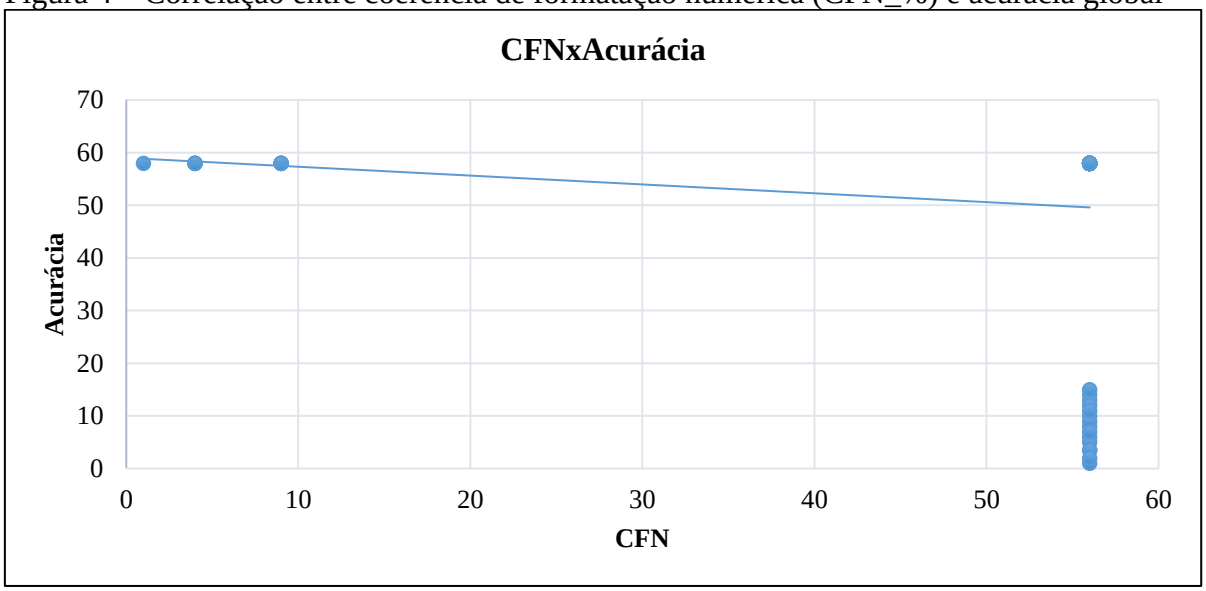
Figura 3 – Correlação entre precisão de rótulos canônicos (PRC_%) e acurácia global



Fonte: elaborado pelo autor.

A Figura 4 ilustra a correlação entre a Coerência de Formatação Numérica (CFN_%) e a acurácia global.

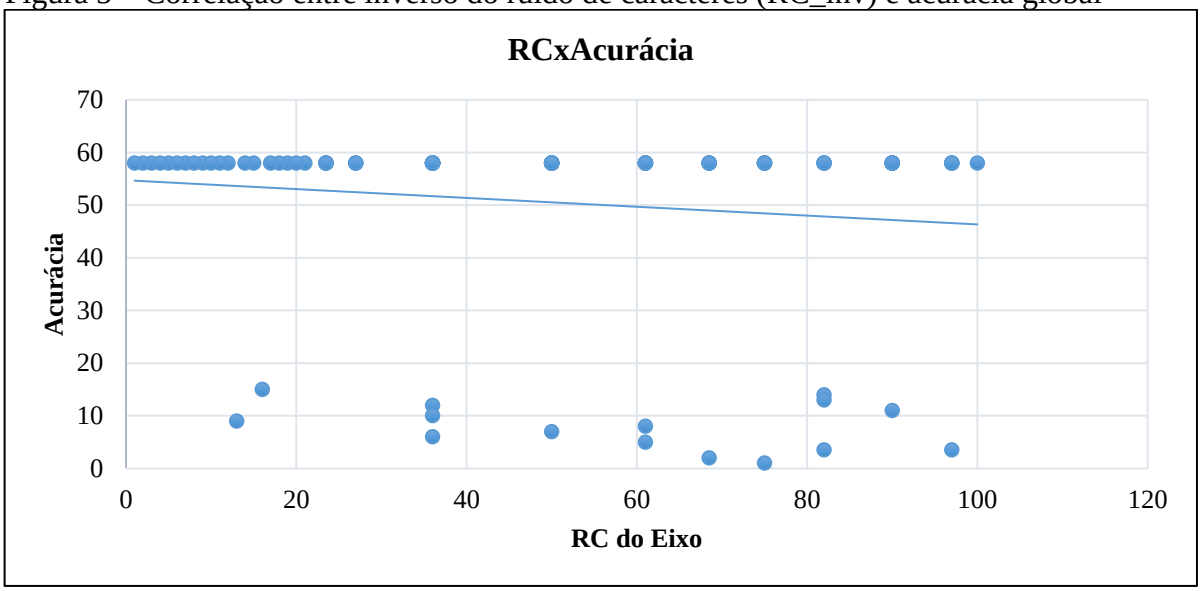
Figura 4 – Correlação entre coerência de formatação numérica (CFN_%) e acurácia global



Fonte: elaborado pelo autor.

A Figura 5 demonstra visualmente a relação entre o Inverso do Ruído de Caracteres (RC_inv) e a acurácia global.

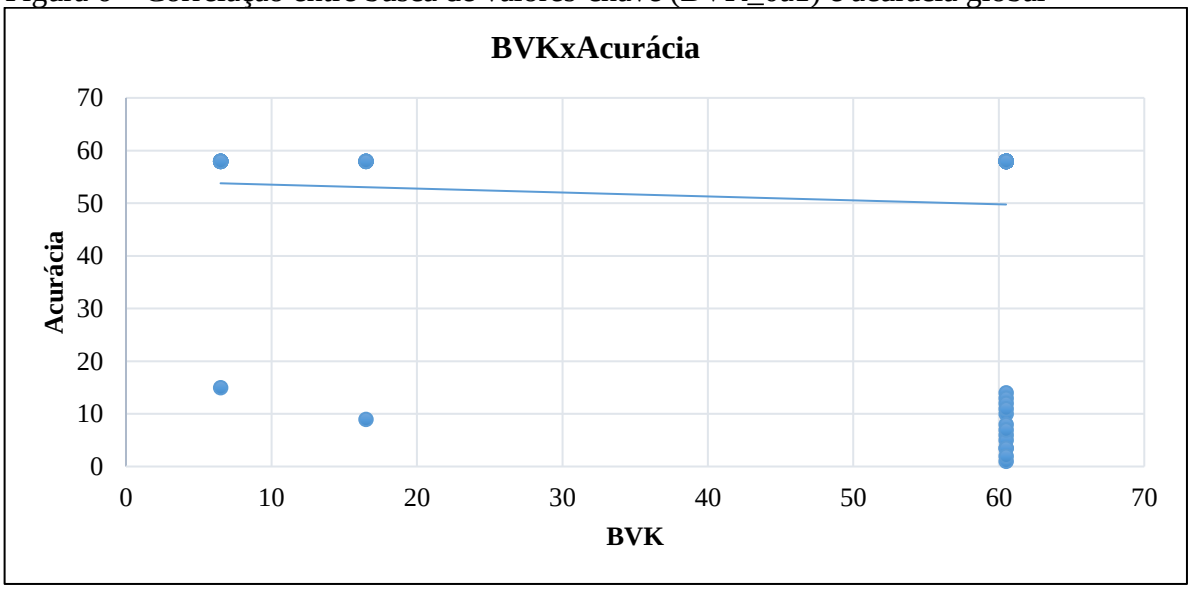
Figura 5 – Correlação entre inverso do ruído de caracteres (RC_inv) e acurácia global



Fonte: elaborado pelo autor.

A Figura 6 exibe a dispersão dos dados ao correlacionar a Busca de Valores-Chave (BVK_0a1) com a acurácia global.

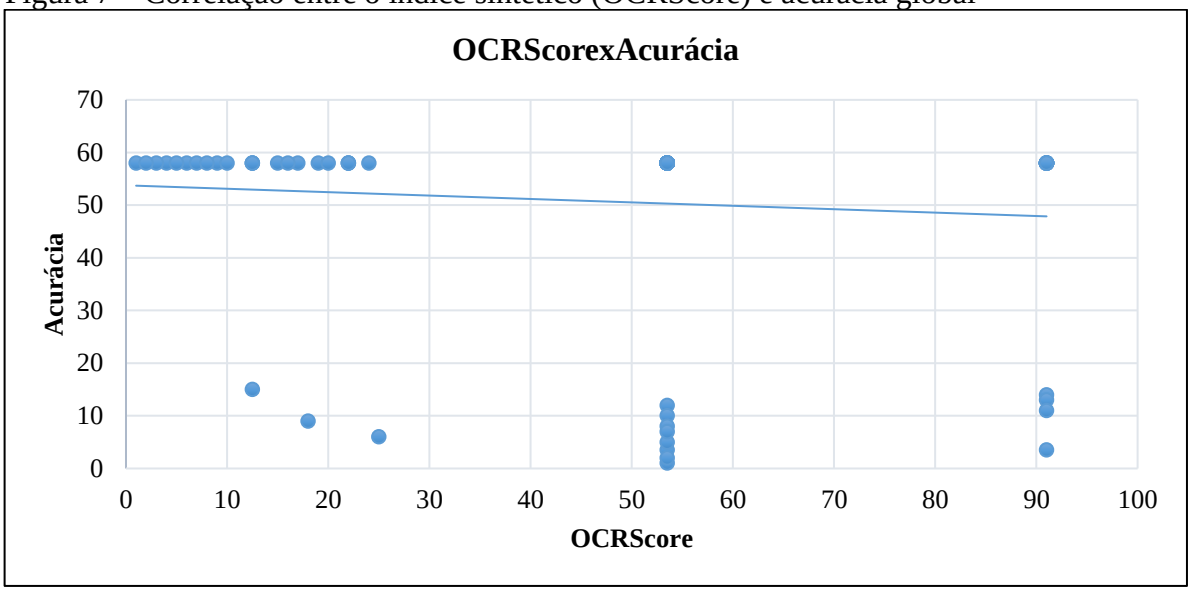
Figura 6 – Correlação entre busca de valores-chave (BVK_0a1) e acurácia global



Fonte: elaborado pelo autor.

A Figura 7 mostra a correlação entre o índice sintético de qualidade (OCRScore) e a acurácia global.

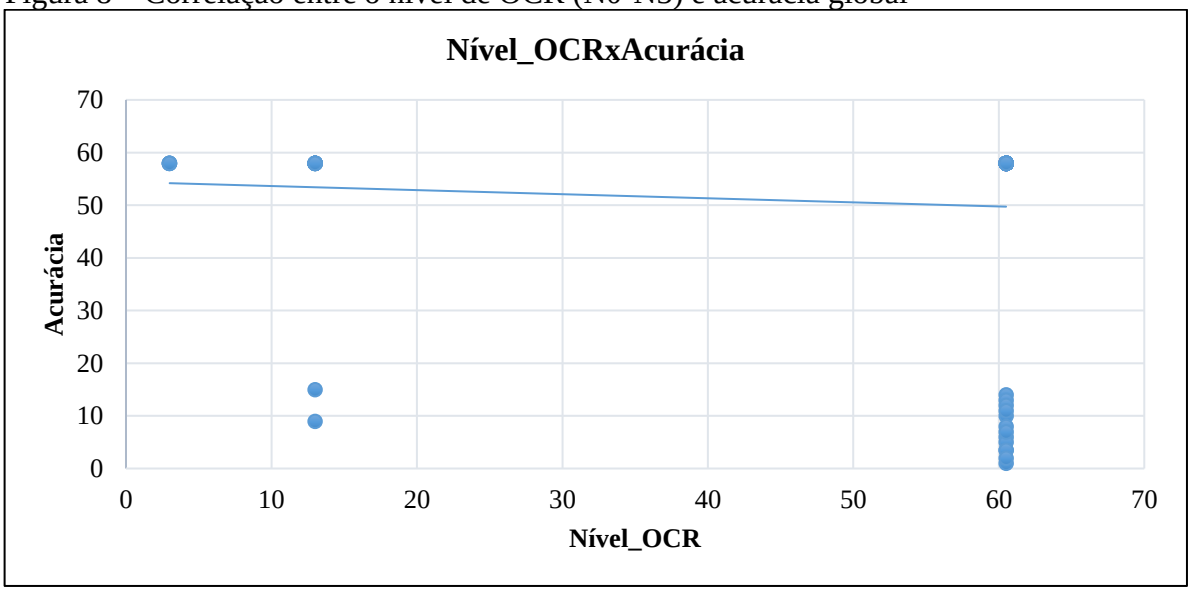
Figura 7 – Correlação entre o índice sintético (OCRScore) e acurácia global



Fonte: elaborado pelo autor.

A Figura 8 apresenta a dispersão da acurácia global em relação ao Nível de OCR (N0-N3) dos documentos.

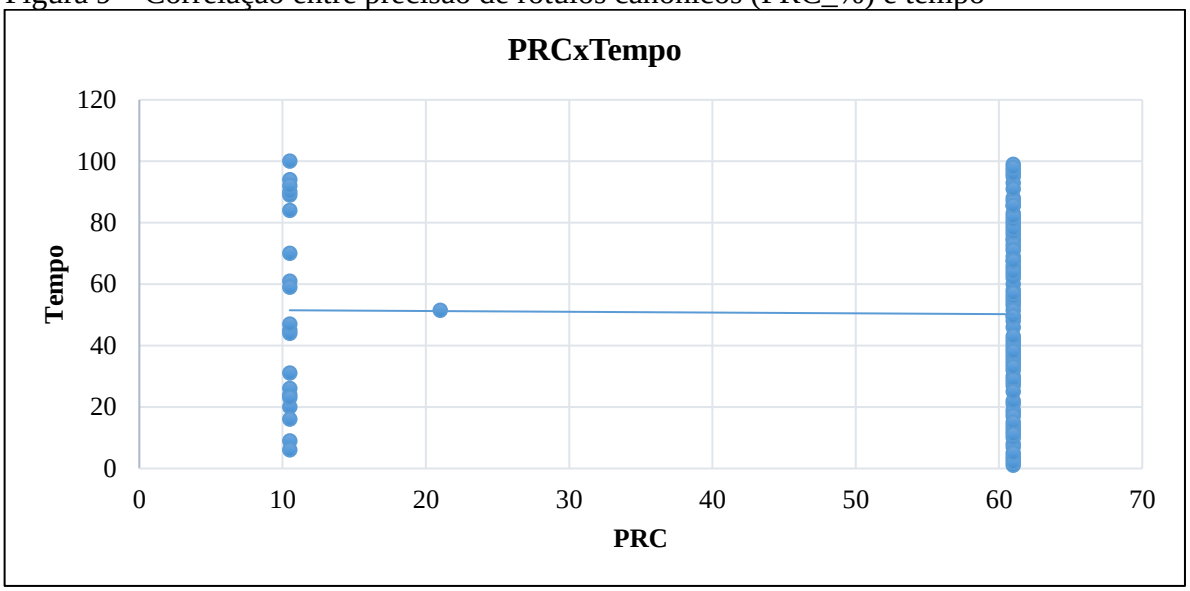
Figura 8 – Correlação entre o nível de OCR (N0-N3) e acurácia global



Fonte: elaborado pelo autor.

Iniciando a análise da eficiência, a Figura 9 correlaciona a Precisão de Rótulos Canônicos (PRC_%) com o tempo de processamento.

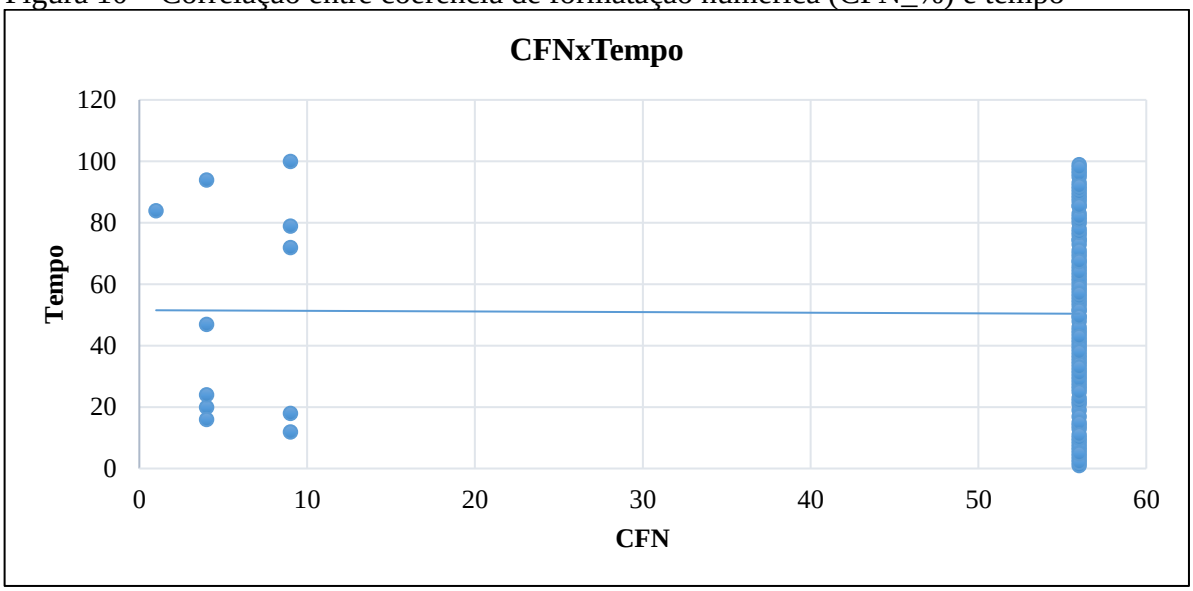
Figura 9 – Correlação entre precisão de rótulos canônicos (PRC_%) e tempo



Fonte: elaborado pelo autor.

A Figura 10 ilustra a relação entre a Coerência de Formatação Numérica (CFN_%) e o tempo de processamento.

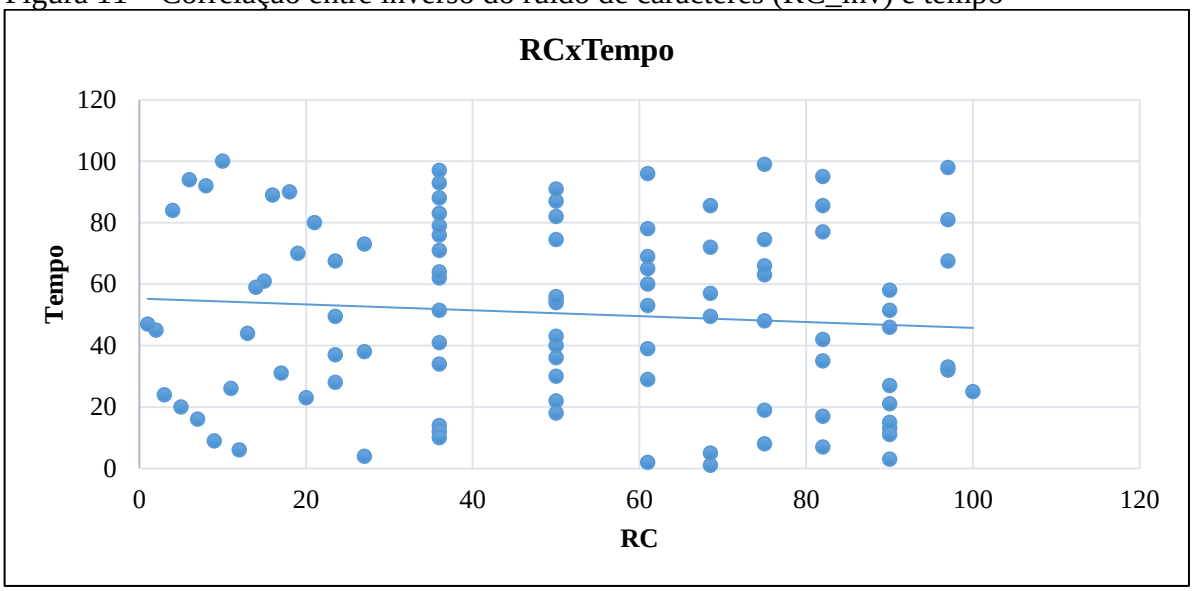
Figura 10 – Correlação entre coerência de formatação numérica (CFN_%) e tempo



Fonte: elaborado pelo autor.

A Figura 11 demonstra a correlação entre o Inverso do Ruído de Caracteres (RC_inv) e o tempo de processamento.

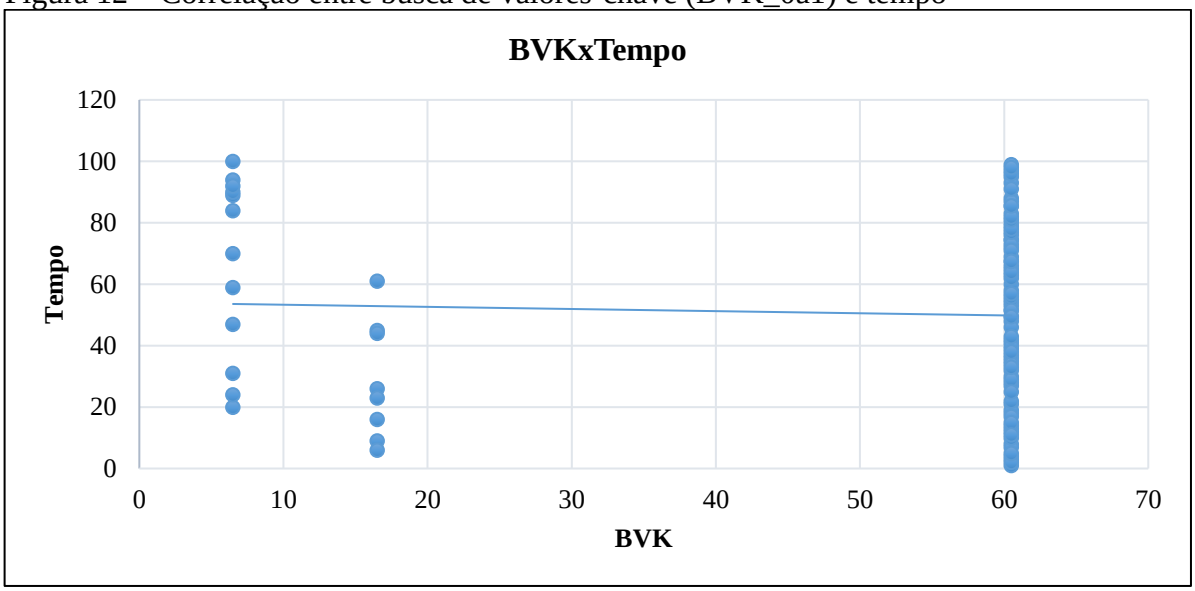
Figura 11 – Correlação entre inverso do ruído de caracteres (RC_inv) e tempo



Fonte: elaborado pelo autor.

A Figura 12 apresenta a dispersão dos dados relacionando a Busca de Valores-Chave (BVK_0a1) ao tempo de processamento.

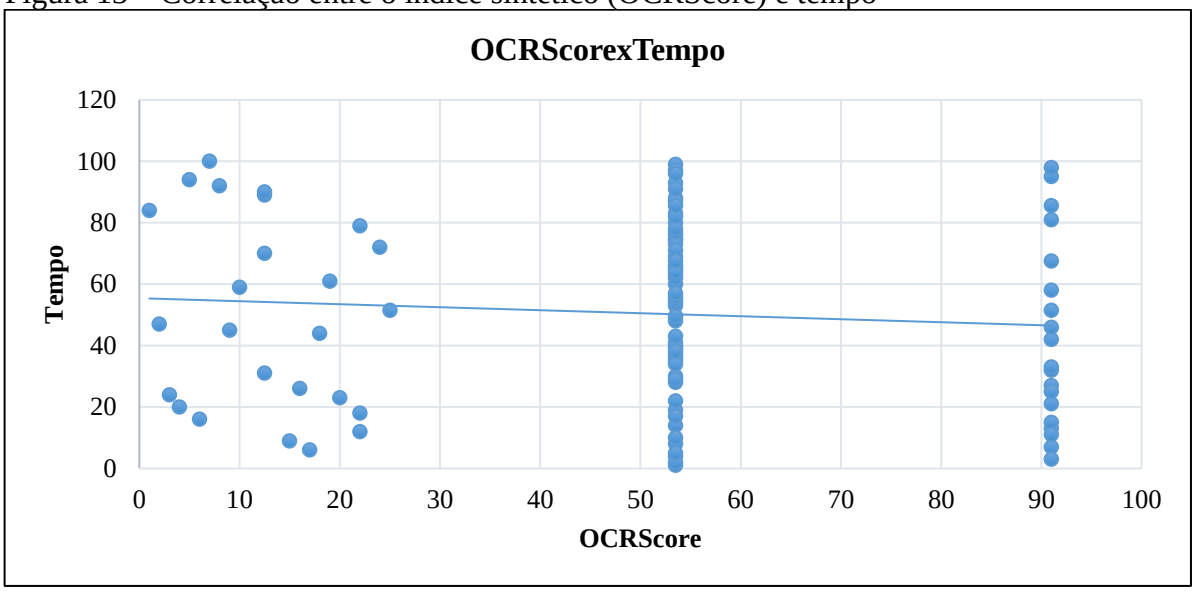
Figura 12 – Correlação entre busca de valores-chave (BVK_0a1) e tempo



Fonte: elaborado pelo autor.

A Figura 13 exibe a correlação entre o índice sintético de qualidade (OCRScore) e o tempo de processamento.

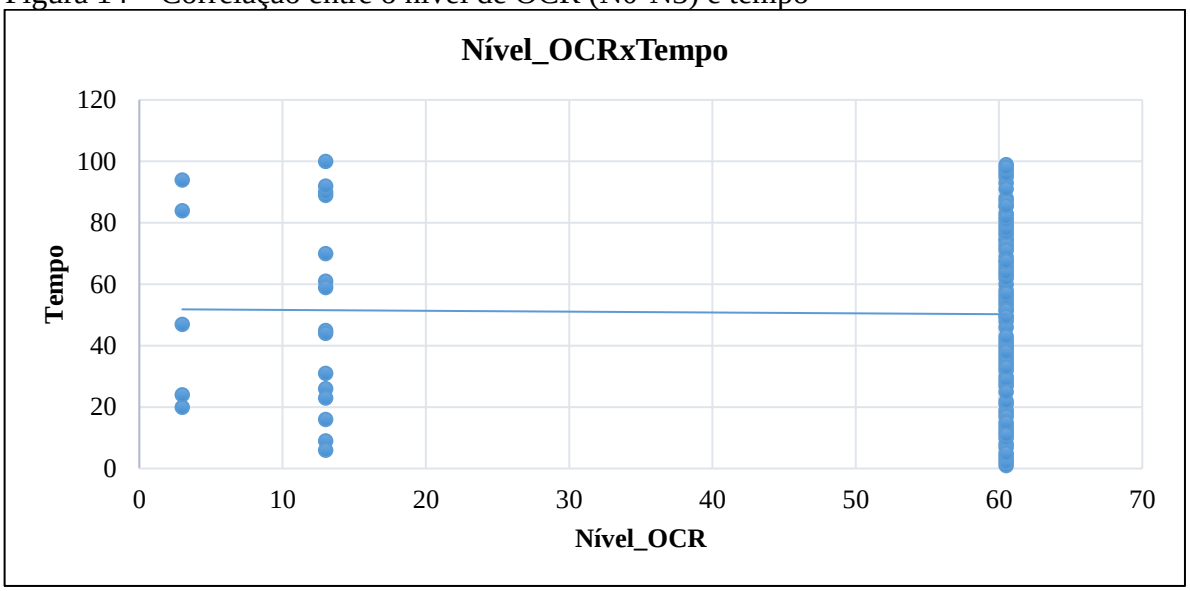
Figura 13 – Correlação entre o índice sintético (OCRScore) e tempo



Fonte: elaborado pelo autor.

Por fim, a Figura 14 mostra a dispersão do tempo de processamento em relação ao Nível de OCR (N0-N3).

Figura 14 – Correlação entre o nível de OCR (N0-N3) e tempo



Fonte: elaborado pelo autor.

A análise dos coeficientes permite tecer duas observações. Primeiramente, no que tange à acurácia, as correlações com todos os indicadores de qualidade documental mostraram-se muito fracas, com valores próximos de zero. O coeficiente mais expressivo foi de -0,147 com

o CFN_%, ainda assim indicando uma associação negligenciável. Este achado sugere que a capacidade do Gemini 2.5 Pro de extrair corretamente as informações foi largamente independente de fatores como a presença de rótulos-chave (PRC_%), a consistência na formatação de números (CFN_%) ou a quantidade de ruído textual (RC_%). Tal resiliência pode ser atribuída à capacidade do modelo de inferir informações a partir do contexto semântico e estrutural do documento, ou ainda, à provável aplicação de técnicas de extração e Reconhecimento Óptico de Caracteres (OCR) internas muito avançadas ao processar o PDF diretamente, permitindo-lhe superar ruídos e inconsistências textuais que seriam barreiras para métodos de extração mais tradicionais.

Em segundo lugar, a relação com a eficiência (Tempo_s) seguiu padrão semelhante. Todos os coeficientes de correlação foram igualmente muito fracos, indicando que o tempo de processamento por contrato não foi significativamente influenciado pela qualidade textual. Deste modo, um documento com baixo OCRScore não demandou, necessariamente, mais tempo para ser processado que um documento de alta qualidade. Isso pode indicar que o principal dispêndio computacional do LLM reside na análise semântica do documento como um todo, uma tarefa de complexidade intrínseca, e não em etapas de pré-processamento ou correção de ruídos textuais.

5 CONCLUSÕES

5.1 DISCUSSÃO DOS RESULTADOS À LUZ DOS OBJETIVOS DA PESQUISA

O objetivo geral do trabalho foi “avaliar o grau de êxito de um modelo de linguagem de larga escala na extração de informações em determinados contratos públicos". Os resultados indicaram um grau de êxito alto. Com oito dos nove campos extraídos atingindo 100% de acurácia e o campo mais complexo (Objeto) superando 91% de acurácia média, o modelo demonstrou uma capacidade de extração de dados com acurácia extremamente elevada para informações estruturadas e é altamente consistente para informações textuais. A variabilidade observada, restrita quase que exclusivamente ao campo “Objeto”, revela que o êxito se mostrou altamente consistente para a maioria dos dados contratuais, sinalizando a maturidade da tecnologia para aplicações práticas.

Concernente ao primeiro objetivo específico, “Identificar a verificabilidade de características e informações dispostas nos contratos", os indicadores de qualidade documental permitiram uma classificação objetiva. A análise demonstrou que a verificabilidade da amostra é predominantemente ótima, com 80% dos contratos classificados no Nível OCR "ótimo" (N3). Ainda que uma parcela de 5% tenha sido classificada como "Fraca" (N1), a ausência de contratos no nível "Crítico" (N0) indica que toda a amostra possuía uma base textual minimamente funcional. Tal diagnóstico quantificou as barreiras técnicas à transparência, evidenciando que, embora a maioria dos documentos possua um padrão elevado, uma minoria ainda apresenta deficiências que poderiam dificultar métodos de automação menos avançados.

O segundo objetivo específico visou “mensurar a acurácia e a eficiência do modelo de linguagem de larga escala na extração automatizada de informações contratuais”. A efetividade, compreendida como a capacidade de gerar resultados precisos em condições reais e não ideais, mostrou-se elevada. Conforme discutido na seção 4.4, a fraca correlação entre a qualidade do OCR e a acurácia foi a principal evidência dessa efetividade. O LLM demonstrou ser uma ferramenta estável, capaz de superar deficiências como a ausência de padronização e o ruído textual, corroborando os achados de Li et al. (2024), que apontaram para o potencial da IA na verificação de informações em documentos governamentais. Pode-se, inclusive, traçar um paralelo direto com os resultados do estudo de Li et al. (2024), que obteve 96% de acurácia e uma redução de 83% no tempo de verificação na avaliação de dados do Diário Oficial do Estado de Santa Catarina. Este trabalho não apenas atingiu um patamar de acurácia similar, como

também o faz em um tempo de processamento médio inferior a 20 segundos por contrato, o que representa um ganho de eficiência de 82,7% em comparação com o método manual. Deste modo, esta pesquisa avançou ao demonstrar que a alta performance é largamente independente da qualidade textual do documento-fonte.

Por fim, o terceiro objetivo específico foi “Analisar a relação entre a qualidade documental dos contratos e o desempenho do modelo, verificando a influência de fatores textuais sobre a extração de dados”. Este objetivo foi alcançado por meio da análise de correlação de Spearman (detalhada na Seção 4.4) que cruzou os indicadores de qualidade textual (como OCRScore, PRC_% e CFN_%) com as métricas de desempenho (acurácia e tempo). A principal conclusão desta análise foi a comprovação da resiliência do modelo: os resultados revelaram coeficientes de correlação muito fracos, próximos de zero, para todas as variáveis. Este achado demonstrou que a alta performance da extração foi largamente independente das deficiências e inconsistências textuais (como ruído de caracteres ou falta de rótulos) presentes nos arquivos-fonte.

5.2 IMPLICAÇÕES, LIMITAÇÕES E RECOMENDAÇÕES

Demonstra-se empiricamente que LLMs de última geração, como o Gemini 2.5 Pro, são capazes de transpor desafios recorrentes no processamento de documentos públicos, notadamente a baixa qualidade do OCR e a falta de padronização. Isso sinaliza uma maturação tecnológica com potencial para aprimorar as práticas de auditoria, transparência e gestão documental no setor público. A automação bem-sucedida de tarefas de extração libera o capital humano para focar em análise crítica, investigação de anomalias e julgamento profissional. Isso sugere uma requalificação necessária do auditor, que se torna menos um "coletor" de dados e mais um "curador" e "analista" do trabalho gerado por IA, o que exige novas competências em ciência de dados, validação de modelos e engenharia de prompt.

Não obstante, esta pesquisa possui limitações que devem ser ponderadas. A amostra, definida por conveniência, restringe-se a 100 contratos de um único departamento (DLOG) e ano (2024), o que limita a generalização dos resultados para outros contextos. Ademais, as métricas de qualidade documental, embora sistematizadas, são *proxies* que não capturam toda a complexidade semântica e visual dos documentos, o que pode ter contribuído para as baixas correlações observadas. Por fim, o estudo concentrou-se em um único modelo de linguagem, e o desempenho pode variar entre diferentes arquiteturas de IA. Acrescenta-se como limitação o

fato de o estudo não ter avaliado a viabilidade econômica da aplicação desta técnica em escala censitária, ou ao menos, significativamente maior. Adicionalmente, o procedimento de extração de dados, embora eficiente (média de 19,53 segundos por contrato), foi realizado de forma manual e sequencial, com o anexo individual de cada PDF em um chat de conversação dedicado, conforme detalhado na Seção 3.2.2. Esta metodologia foi intencionalmente adotada para garantir a medição precisa do tempo de processamento por contrato, atendendo ao segundo objetivo específico da pesquisa, que era mensurar a eficiência. No entanto, para fins de aplicação em larga escala por órgãos públicos (análise censitária), recomenda-se o uso de APIs (Application Programming Interfaces). O uso de uma API permitiria a automação do upload e processamento dos contratos em bloco, eliminando a necessidade da intervenção humana para cada documento e reduzindo o tempo total da tarefa de extração.

Embora o tempo de processamento por documento seja baixo, o custo computacional para processar milhões de contratos pode representar uma barreira prática significativa para a adoção em larga escala por órgãos públicos, algo que não foi objeto desta análise. Outra limitação inerente ao uso de LLMs de ponta é a questão da "explicabilidade". Embora a acurácia de saída tenha sido validada, o processo de inferência (como o modelo identificou e decidiu qual era o dado correto) permanece uma "caixa-preta". Em um contexto de auditoria formal, a incapacidade de rastrear o "raciocínio" da ferramenta pode ser um obstáculo para sua maior aceitação como prova de auditoria.

A aplicação do LLM no contexto desta pesquisa possui implicações diretas para a Auditoria Governamental e o Controle Interno. O nível de acurácia da extração assegura a confiabilidade dos dados obtidos, isto permite afirmar que as informações geradas podem servir como evidência de auditoria válida para procedimentos de compliance e análise de risco. A redução de 82,7% no tempo de processamento por contrato pode alterar o escopo da fiscalização. Esta eficiência possibilita a transição do controle baseado em amostragem para a análise censitária, o que mitiga o risco de auditoria associado à não detecção de desvios em itens não amostrados. Assim, a acurácia e a velocidade de processamento verificadas são fatores tecnológicos que podem servir de suporte à implementação da Auditoria Contínua, sendo possível o monitoramento proativo de contratos no setor público.

Com base no exposto, derivam-se recomendações. Aos órgãos públicos, recomenda-se a adoção de políticas de padronização na elaboração de contratos e a priorização da publicação de documentos em formatos nativo-digitais, em vez de digitalizados, para melhorar a fidedignidade de qualquer ferramenta de extração automatizada. Para pesquisas futuras, sugere-se a ampliação da amostra, a inclusão de diferentes tipos de documentos públicos e a

realização de análises comparativas entre distintos LLMs. Adicionalmente, estudos futuros poderiam investigar quais características estruturais e de *layout* dos documentos, para além da qualidade textual, exercem maior impacto no desempenho dos modelos.

Em suma, esta pesquisa avança em relação a uma simples validação tecnológica. Ela oferece uma evidência empírica de que é possível superar desafios recorrentes entre dados "presos" em documentos e a efetiva criação de dados estruturados e auditáveis. O estudo demonstra que as tecnologias de linguagem atuais já permitem um controle mais detalhado e eficiente, mostrando-se capazes de lidar com as inconsistências dos documentos reais. O desafio que se impõe, portanto, migra do campo técnico para o estratégico: a adoção pelas instituições, a integração de sistemas e a capacitação dos agentes públicos para operar neste novo cenário de transparência ampliada pela tecnologia.

REFERÊNCIAS

BATAGLIA, Murilo Borsio; FARRANHA, Ana Claudia. Controle social e acesso à informação: o papel da transparência passiva no enfrentamento à corrupção. **Interfaces Científicas - Direito**, [S. l.], v. 6, n. 3, p. 27–42, 2018. DOI: 10.17564/2316-381X.2018v6n3p27–42. Disponível em: https://periodicos.set.edu.br/direito/article/view/5865. Acesso em: 26 abr. 2025.

BRASIL. [Constituição (1988)]. **Constituição da República Federativa do Brasil de 1988**. Brasília, DF: Presidência da República, [1988]. Disponível em: https://www.planalto.gov.br/ccivil_03/constituicao/constituicao.htm. Acesso em: 25 maio 2025.

BRASIL. **Decreto nº 10.278, de 18 de março de 2020**. Regulamenta o disposto no inciso X do caput do art. 3º da Lei nº 13.874, de 20 de setembro de 2019, para dispor sobre a digitalização de documentos públicos ou privados e a sua produção e arquivamento em meio eletrônico. Diário Oficial da União: seção 1, Brasília, DF, p. 6, 19 mar. 2020. Disponível em: https://www.planalto.gov.br/ccivil_03/_Ato2019-2022/2020/Decreto/D10278.htm. Acesso em: 26 abr. 2025.

BRASIL. **Lei nº 12.527, de 18 de novembro de 2011**. Regula o acesso a informações previsto na Constituição. Diário Oficial da União: seção 1, Brasília, DF, 18 nov. 2011. Disponível em: https://www.planalto.gov.br/ccivil_03/_ato2011-2014/2011/lei/l12527.htm. Acesso em: 25 maio 2025.

BRASIL. **Lei nº 13.460, de 26 de junho de 2017**. Dispõe sobre participação, proteção e defesa dos direitos do usuário dos serviços públicos da administração pública. Diário Oficial da União: seção 1, Brasília, DF, 26 jun. 2017. Disponível em: https://www.planalto.gov.br/ccivil_03/_Ato2015-2018/2017/Lei/L13460.htm. Acesso em: 25 maio 2025.

BRASIL. **Lei nº 14.129, de 29 de março de 2021**. Dispõe sobre princípios, regras e instrumentos para o Governo Digital e para o aumento da eficiência pública. Diário Oficial da União: seção 1, Brasília, DF, 30 mar. 2021. Disponível em: https://www.planalto.gov.br/ccivil_03/_Ato2019-2022/2021/Lei/L14129.htm. Acesso em: 25 maio 2025.

BRASIL. **Lei nº 14.133, de 1º de abril de 2021**. Institui a nova Lei de Licitações e Contratos Administrativos. Diário Oficial da União: seção 1, Brasília, DF, 1 abr. 2021. Disponível em: https://www.planalto.gov.br/ccivil_03/_Ato2019-2022/2021/Lei/L14133.htm. Acesso em: 25 maio 2025.

BRASIL. Ministério da Saúde. **Contratos do Departamento de Logística em Saúde – DLOG**. Disponível em: https://www.gov.br/saude/pt-br/acesso-a-informacao/licitacoes-e-contratos/contratos-dlog. Acesso em: 26 abr. 2025.

CARVALHO FILHO, José dos Santos. **Manual de Direito Administrativo**. 39. ed. São Paulo: Atlas, 2025.

COELHO, T. R. et al. Transparência governamental nos estados e grandes municípios brasileiros: uma “dança dos sete véus” incompleta? **Cadernos Gestão Pública e Cidadania**, São Paulo, v. 23, n. 75, 2018. DOI: 10.12660/cgpc.v23n75.73447. Disponível em: https://periodicos.fgv.br/cgpc/article/view/73447. Acesso em: 26 abr. 2025.

CRESWELL, John W.; CLARK, Vicki L. Plano. **Designing and conducting mixed methods research**. 3. ed. Los Angeles: SAGE, 2017.

CRESWELL, John W.; CRESWELL, J. David. **Research design: qualitative, quantitative, and mixed methods approaches**. 6. ed. Thousand Oaks: SAGE Publications, 2022.

DAIR.AI. **Guia de Engenharia Prompt - Elementos de um prompt**. [S.l.]: DAIR.AI,. Disponível em: promptingguide.ai/pt/introduction/elements. Acesso em: 19 out. 2025.

DI PIETRO, Maria Sylvia Zanella. **Direito Administrativo**, 38. ed., São Paulo: Editora Forense, 2025.

FADGI - Federal Agencies Digital Guidelines Initiative. **Technical Guidelines for Digitizing Cultural Heritage Materials**. 3. ed., 01 maio 2023. Disponível em: https://www.digitizationguidelines.gov/guidelines/FADGI%20Technical%20Guidelines%20for%20Digitizing%20Cultural%20Heritage%20Materials_3rd%20Edition_05092023.pdf. Acesso em: 19 out. 2025.

GIL, Antonio Carlos. **Como elaborar projetos de pesquisa**. 7. ed. São Paulo: Atlas, 2022.

GIL, Antônio Carlos. **Métodos e técnicas de pesquisa social**. 6. ed. São Paulo: Atlas, 2008.

GRIMMER, Justin; STEWART, Brandon M. Text as data: the promise and pitfalls of automatic content analysis methods for political texts. **Political Analysis**, v. 21, n. 3, p. 267–297, 2013. DOI: 10.1093/pan/mps028. Disponível em: https://www.cambridge.org/core/journals/political-analysis/article/text-as-data-the-promise-and-pitfalls-of-automatic-content-analysis-methods-for-political-texts/F7AAC8B2909441603FEB25C156448F20. Acesso em: 19 out. 2025.

HAMDI, Ahmed et al. In-depth analysis of the impact of OCR errors on named entity recognition and linking. **Natural Language Engineering**, v. 29, n. 2, p. 425–448, 2023. DOI: 10.1017/S1351324922000110. Disponível em: https://www.cambridge.org/core/journals/natural-language-engineering/article/abs/indepth-analysis-of-the-impact-of-ocr-errors-on-named-entity-recognition-and-linking/C732399FF72BAFE8FF830BB1F5ED7576. Acesso em: 19 out. 2025.

JANSSEN, Marijn; CHARALABIDIS, Yannis; ZUIDERWIJK, Anneke. Benefits, adoption barriers and myths of open data and open government. **Information Systems Management**, v. 29, n. 4, p. 258–268, 2012. DOI: 10.1080/10580530.2012.716740. Disponível em: https://www.tandfonline.com/doi/abs/10.1080/10580530.2012.716740. Acesso em: 19 out. 2025.

KOKINA, Julia; DAVENPORT, Thomas H. The Emergence of Artificial Intelligence: How Automation is Changing Auditing. **Journal of Emerging Technologies in Accounting**, v. 14, n. 1, p. 115-122, 2017. Disponível em: https://doi.org/10.2308/jeta-51730. Acesso em: 6 jul. 2025.

LI, Huaxia et al. **Enhancing Continuous Auditing with Large Language Models: AI-Assisted Real-Time Accounting Information Cross-verification**. SSRN, 2024. Disponível em: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4692960. Acesso em: 26 abr. 2025.

MASLEJ, Nestor et al. **The AI Index 2025 Annual Report**. Stanford, CA: AI Index Steering Committee, Institute for Human-Centered AI, Stanford University, 2025. Disponível em: https://hai.stanford.edu/ai-index. Acesso em: 26 abr. 2025.

MELLO, Celso Antônio Bandeira de. **Curso de Direito Administrativo**. 40. ed. São Paulo: Malheiros, 2023.

OECD. **Governing with Artificial Intelligence**. Paris: OECD Publishing, 2025. Disponível em: https://www.oecd.org/en/publications/2025/06/governing-with-artificial-intelligence_398fa287/full-report/ai-in-public-procurement_2e095543.html. Acesso em: 19 out. 2025.

OECD. **Managing risks in the public procurement of goods, services and infrastructure**. Paris: OECD Publishing, 2023. Disponível em: https://www.oecd.org/en/publications/managing-risks-in-the-public-procurement-of-goods-services-and-infrastructure_45667d2f-en.html. Acesso em: 19 out. 2025.

SCHMIDT, Lars; MAIDEN, Jeffrey; MAIDEN, Neil. Fiscal data in text: Information extraction from audit reports using Natural Language Processing. **Data & Policy**, Cambridge University Press, v. 4, e18, 2023. Disponível em: https://www.cambridge.org/core/journals/data-and-policy/article/fiscal-data-in-text-information-extraction-from-audit-reports-using-natural-language-processing/F4CAA159BD8C5C71873D85FCF1E4AA96. Acesso em: 26 abr. 2025.

SOYLU, Ahmet et al. Data quality barriers for transparency in public procurement. **Information**, v. 13, n. 2, art. 99, 2022. DOI: 10.3390/info13020099. Disponível em: https://www.mdpi.com/2078-2489/13/2/99. Acesso em: 19 out. 2025.

VASWANI, Ashish et al. Attention Is All You Need. **Advances in Neural Information Processing Systems (NeurIPS)**, 2017, [S.l.]. Anais [...]. [S.l.: s.n.], 2017. Disponível em: https://papers.neurips.cc/paper/7181-attention-is-all-you-need.pdf. Acesso em: 25 mai. 2025.

XU, Derong et al. Large Language Models for Generative Information Extraction: A Survey. **Frontiers of Computer Science**, [S.l.: s.n.], 2024. DOI: 10.1007/s11704-024-40555-y. Disponível em: https://arxiv.org/abs/2312.17617. Acesso em: 19 out. 2025.

ZHAO, Wayne Xin et al. **A Survey of Large Language Models**. [S.l.: s.n.], 2023. Disponível em: https://arxiv.org/pdf/2303.18223. Acesso em: 26 abr. 2025.

APÊNDICE A – CÓDIGO EM PYTHON UTILIZADO

```
# # ============= BLOCO 1 — INSTALAÇÃO DAS BIBLIOTECAS =============
!pip -q install pymupdf pdfplumber pandas openpyxl
from pathlib import Path
Path("/content/contratos").mkdir(parents=True, exist_ok=True)  # subir os PDFs aqui
print("Pasta para os contratos:", "/content/contratos")

# # ==================== BLOCO 2 — CONFIGURAÇÕES ====================

import re, unicodedata
from pathlib import Path
from typing import List, Dict
from glob import glob
import pandas as pd

# -------- Config --------
ROTULOS = ["CNPJ", "Objeto", "Valor Total", "Vigência", "Elemento de Despesa"]
REGEX_CNPJ = re.compile(r"\b\d{2}\.?\d{3}\.?\d{3}/?\d{4}-?\d{2}\b")
REGEX_MOEDA = re.compile(r"\b\d{1,3}(?:\.\d{3})*,\d{2}\b")
SAMPLE_CHARS_PER_PAGE = 1500

# -------- Backends de PDF (PyMuPDF priorizado; fallback pdfplumber) --------
def _load_pymupdf():
    try:
        import fitz  # PyMuPDF
        return fitz
    except Exception:
        return None

def _load_pdfplumber():
    try:
        import pdfplumber
        return pdfplumber
    except Exception:
        return None

FITZ = _load_pymupdf()
PDFPLUMBER = _load_pdfplumber()

def extract_doc_texts(pdf_path: Path) -> List[str]:
    texts = []
    if FITZ is not None:
        try:
            doc = FITZ.open(pdf_path)
            page_count = doc.page_count
            pages_to_scan = range(page_count)
            for i in pages_to_scan:
```

```
                page = doc.load_page(i)
                t = page.get_text("text") or ""
                texts.append(t)
            doc.close()
            return texts
        except Exception:
            pass
    if PDFPLUMBER is not None:
        try:
            with PDFPLUMBER.open(pdf_path) as pdf:
                page_count = len(pdf.pages)
                pages_to_scan = range(page_count)
                for i in pages_to_scan:
                    t = pdf.pages[i].extract_text() or ""
                    texts.append(t)
            return texts
        except Exception:
            pass
    return texts  # pode vir vazio se não conseguiu ler

# -------- Métricas --------
def pct(a: int, b: int) -> float:
    return 0.0 if b == 0 else (100.0 * a / b)

def prc_percent(full_text: str) -> float:
    hits = sum(1 for r in ROTULOS if r.lower() in full_text.lower())
    return pct(hits, len(ROTULOS))

def cfn_percent(samples: List[str]) -> float:
    ok = 0
    for s in samples:
        if REGEX_MOEDA.search(s):
            ok += 1
        else:
            ok += 1 if re.search(r"\d{3,}[\.,]?\d*", s) else 0
    return pct(ok, len(samples)) if samples else 0.0

def rc_percent(samples: List[str]) -> float:
    bad = 0; total = 0
    for s in samples:
        for ch in s:
            total += 1
            cat = unicodedata.category(ch)
            if ch == "\ufffd" or cat.startswith("C"):
                bad += 1
    return 0.0 if total == 0 else (100.0 * bad / total)

def bvk_score(full_text: str) -> float:
    has_cnpj = bool(REGEX_CNPJ.search(full_text))
    has_valor = bool(REGEX_MOEDA.search(full_text))
```

```
    return 1.0 if (has_cnpj and has_valor) else (0.5 if (has_cnpj or has_valor) else 0.0)

def table_heuristic(full_text: str) -> int:
    lines = full_text.splitlines()
    score = 0
    for ln in lines:
        if "\t" in ln:
            score += 1
        elif re.search(r"(  ){2,}", ln):
            score += 1
        elif re.search(r"(;|,).*(;|,).*(;|,)", ln):
            score += 1
    return 1 if score >= max(5, int(0.02 * len(lines))) else 0

def sample_pages(texts: List[str]) -> List[str]:
    if not texts: return []
    n = len(texts)
    idx = [0] if n == 1 else ([0, 1] if n == 2 else [0, n//2, n-1])
    return [texts[i][:SAMPLE_CHARS_PER_PAGE] for i in idx]

def infer_level(ocr_score: float) -> str:
    if ocr_score < 25: return "N0"
    if ocr_score < 50: return "N1"
    if ocr_score < 80: return "N2"
    return "N3"

# -------- Processar 1 PDF --------
def process_single(pdf_path: str | Path) -> Dict[str, object]:
    pdf_path = Path(pdf_path)
    texts = extract_doc_texts(pdf_path)
    n_pages = len(texts)
    pages_with_text = sum(1 for t in texts if t and t.strip())

    COP = pct(pages_with_text, n_pages)
    full_text = "\n".join(texts)
    PRC = prc_percent(full_text)

    samples = sample_pages(texts)
    CFN = cfn_percent(samples)
    RC = rc_percent(samples)

    ITB = table_heuristic(full_text)
    BVK = bvk_score(full_text)

    OCRScore = (0.20 * COP +
                0.25 * PRC +
                0.20 * CFN +
                0.10 * (100 - RC) +
                0.15 * (ITB * 100) +
                0.10 * (BVK * 100))
```

```
    nivel = infer_level(OCRScore)

    return {
        "arquivo_pdf": pdf_path.name,
        "paginas_totais": n_pages,
        "paginas_com_texto": pages_with_text,
        "COP_%": round(COP, 1),
        "PRC_%": round(PRC, 1),
        "CFN_%": round(CFN, 1),
        "RC_%": round(RC, 1),
        "ITB_0ou1": int(ITB),
        "BVK_0a1": float(BVK),
        "OCRScore_0a100": round(OCRScore, 1),
        "Nivel_OCR_(N0-N3)": nivel
    }

# -------- Rodar todos e salvar tabela --------
def run_many(contracts: List[str | Path],
             out_csv: str = "AQ_OCR_resultados.csv",
             out_xlsx: str = "AQ_OCR_resultados.xlsx") -> pd.DataFrame:
    rows = []
    errors = []
    for p in contracts:
        p = str(p)
        try:
            rows.append(process_single(p))
        except Exception as e:
            errors.append((p, repr(e)))
    df = pd.DataFrame(rows)
    df.to_csv(out_csv, index=False, encoding="utf-8-sig")
    try:
        import openpyxl  # noqa
        df.to_excel(out_xlsx, index=False, sheet_name="AQ_OCR")
    except Exception:
        pass
    return df


# ================== BLOCO 3 — RODAR CONTRATOS ==================
# Ordena, roda, salva e mostra 100 linhas

def natsort_key(s: str):
    return [int(t) if t.isdigit() else t.lower() for t in re.split(r'(\d+)', s)]

# 1) coletar contratos
collected = glob("/content/contratos/**/*", recursive=True)
contracts = [p for p in collected if p.lower().endswith(".pdf")]
contracts = sorted(set(contracts), key=natsort_key)

# 2) rodar e gerar tabela
```

```
df = run_many(
    contracts,
    out_csv="AQ_OCR_resultados.csv",
    out_xlsx="AQ_OCR_resultados.xlsx"
)

# 3) exibir 100 linhas na saída do notebook
pd.set_option("display.max_rows", 100)
display(df.head(100))
```