

Título: APLICAÇÃO DE LLM NO CONTROLE DE CONFORMIDADE EM LICITAÇÕES DE TECNOLOGIA DA INFORMAÇÃO

Autor: Leonardo Alamy Martins

Orientador: Aloísio Dourado Neto

Coletânea de Pós-Graduação
**Controle Governamental:
Tecnologias para Inovação**



REPÚBLICA FEDERATIVA DO BRASIL
TRIBUNAL DE CONTAS DA UNIÃO

MINISTROS

Vital do Rêgo Filho (Presidente)
Jorge Antônio de Oliveira Francisco (Vice-Presidente)
Walton Alencar Rodrigues
Benjamin Zymler
João Augusto Ribeiro Nardes
Aroldo Cedraz de Oliveira
Bruno Dantas
Antonio Augusto Junho Anastasia
Jhonatan de Jesus

MINISTROS-SUBSTITUTOS

Augusto Sherman Cavalcanti
Marcos Bemquerer Costa
Weder de Oliveira

MINISTÉRIO PÚBLICO JUNTO AO TCU

Cristina Machado da Costa e Silva (Procuradora-Geral)
Lucas Furtado (Subprocurador-Geral)
Paulo Soares Bugarin (Subprocurador-Geral)
Marinus Eduardo de Vries Marsico (Procurador)
Júlio Marcelo de Oliveira (Procurador)
Sérgio Ricardo Costa Caribé (Procurador)
Rodrigo Medeiros de Lima (Procurador)



DIRETOR-GERAL

Adriano Cesar Ferreira Amorim

**DIRETORA DE RELAÇÕES INSTITUCIONAIS,
PÓS-GRADUAÇÃO E PESQUISAS**

Flávia Lacerda Franco Melo Oliveira

CONSELHO ACADÊMICO

Junnius Marques Arifa

André Anderson de Oliveira Barbosa

Edans Flávius de Oliveira Sandes

Alberto de Sousa Rocha Júnior

Rafael Silveira e Silva

Pedro Paulo de Moraes

COORDENADOR ACADÊMICO

Edans Flávius de Oliveira Sandes

COORDENADORA PEDAGÓGICA

Marta Eliane Silveira da Costa Bissacot

COORDENADORA EXECUTIVA

Maria das Graças da Silva Duarte de Abreu

PROJETO GRÁFICO E CAPA

Núcleo de Comunicação – NCOM/ISC

APLICAÇÃO DE LLM NO CONTROLE DE CONFORMIDADE EM LICITAÇÕES DE TECNOLOGIA DA INFORMAÇÃO

Autor: Leonardo Alamy Martins

Artigo de conclusão de curso submetido ao Instituto Serzedello Corrêa do Tribunal de Contas da União como requisito parcial para a obtenção do grau de especialista.

Orientador(a):

Prof. Aloísio Dourado Neto

Banca examinadora:

Prof. Eric Hans Messias da Silva

REFERÊNCIA BIBLIOGRÁFICA

ALAMY MARTINS, Leonardo. **APLICAÇÃO DE LLM NO CONTROLE DE CONFORMIDADE EM LICITAÇÕES DE TECNOLOGIA DA INFORMAÇÃO**. 2024. Monografia (Especialização em Avaliação de Políticas Públicas) – Instituto Serzedello Corrêa, Escola Superior do Tribunal de Contas da União, Brasília DF. 58 fl.

CESSÃO DE DIREITOS

NOME DO(A) AUTOR(A): Leonardo Alamy Martins
TÍTULO: APLICAÇÃO DE LLM NO CONTROLE DE CONFORMIDADE EM LICITAÇÕES DE TECNOLOGIA DA INFORMAÇÃO
GRAU/ANO: Especialista/2025

É concedido ao Instituto Serzedello Corrêa (ISC) permissão para reproduzir cópias deste Trabalho de Conclusão de Curso e emprestar ou vender tais cópias somente para propósitos acadêmicos e científicos. Do mesmo modo, o ISC tem permissão para divulgar este documento em biblioteca virtual, em formato que permita o acesso via redes de comunicação e a reprodução de cópias, desde que protegida a integridade do conteúdo dessas cópias e proibido o acesso a partes isoladas desse conteúdo. O autor reserva outros direitos de publicação e nenhuma parte deste documento pode ser reproduzida sem a autorização por escrito do autor.

Nome
e-mail

FICHA CATALOGRÁFICA

L131a Alamy Martins, Leonardo

APLICAÇÃO DE LLM NO CONTROLE DE CONFORMIDADE
EM LICITAÇÕES DE TECNOLOGIA DA INFORMAÇÃO / Autor. –
Brasília: ISC/TCU, 2021.
58 fl. (Artigo de Especialização)

1. Controle Governamental: Tecnologias para Inovação. 2. LLM. 3.
Controle de Conformidade. I. Título.

CDU 02
CDD 020

APLICAÇÃO DE LLM NO CONTROLE DE CONFORMIDADE EM LICITAÇÕES DE TECNOLOGIA DA INFORMAÇÃO

Leonardo Alamy Martins

Trabalho de conclusão do curso de pós-graduação *lato sensu* em Controle Governamental: Tecnologias para Inovação realizado pelo Instituto Serzedello Corrêa como requisito para a obtenção do título de especialista.

Brasília, 11 de dezembro de 2024.

Banca Examinadora:

Prof. Aloísio Dourado Neto
Orientador
Tribunal de Contas da União

Prof. Eric Hans Messias da Silva
Avaliador
Tribunal de Contas da União

Dedico esse trabalho à equipe de Auditoria de Tecnologia da Informação da Controladoria-Geral da União.

Agradecimentos

À minha esposa, Fernanda, e aos meus filhos, Henrique e Davi, por tornarem minha vida cada mais feliz e por darem sentido a ela; pela paciência, pelo suporte e compreensão que me concederam nos últimos dois anos para que fosse possível chegar até aqui e concluir mais essa importante etapa.

Aos meus pais e à minha falecida avó, dona Dinoca (in-memoriam), pela formação que me propiciaram, viabilizando que eu tivesse condições de chegar até aqui.

À Controladoria-Geral da União, por, mais uma vez, conceder a mim a oportunidade de me capacitar, participando de um curso tão relevante para o qual produzo este trabalho.

Ao ISC/TCU, pela qualidade e pela flexibilidade de seu programa de Pós-graduação em Controle Governamental.

Aos meus colegas de turma, pelo apoio durante as aulas e na realização dos trabalhos elaborados ao longo das disciplinas desta Pós-Graduação.

Ao meu orientador, professor doutor Aloísio Dourado Neto, pela atenção, pela paciência e pelas contribuições extremamente valiosas que guiaram a elaboração deste TCC.

Resumo

Este trabalho avalia o potencial de uso de Grandes Modelos de linguagem para ampliar a eficiência de trabalhos de auditoria preventiva em contratações públicas. Utilizando o modelo GPT-4o, o estudo avalia a conformidade de estudos técnicos preliminares de contratações de soluções de tecnologia da informação frente aos requisitos normativos da Instrução Normativa SGD/ME nº 94/2022 e compara as respostas fornecidas pelo LLM com as análises feitas por auditores internos governamentais. Os resultados demonstram que a análise automatizada tem bom desempenho na identificação de não conformidades, bem como realiza análises bem fundamentadas levando menos de um minuto para cada avaliar cada ETP, configurando-se em uma ferramenta poderosa para ampliação da produtividade e da eficiência dos trabalhos de auditoria. O trabalho destaca, ainda, que a ferramenta implementada com IA tende a concluir pela conformidade parcial em casos que os auditores opinam pela conformidade total, sobretudo devido à extensão dos requisitos normativos e à dificuldade de interpretação contextualizada da Instrução normativa e dos Estudos Técnicos Preliminares. Entre as principais limitações encontradas, destacam-se (1) a inexistência de dados públicos de avaliações de conformidade normativa realizada por órgãos de auditoria; (2) a consequente necessidade de utilização de avaliações humanas realizadas por auditores de forma individual, e não institucional; (3) a quantidade relativamente baixa de ETPs analisados por auditores para fins de comparação com a solução automatizada; (4) a utilização de um único LLM; e (5) o fato de que, nas instruções para o modelo de linguagem, terem sido passadas apenas alguns dos trechos mais relevantes dos documentos analisados, e não o conteúdo completo deles. Os resultados deste trabalho podem beneficiar equipes de auditoria tanto no nível federal quanto nas esferas estadual e municipal, além de terem potencial de aplicação para qualquer tipo de objeto auditado, e não apenas para Tecnologia da Informação. Por fim, a solução baseada em inteligência artificial proposta neste estudo pode ser empregada também diretamente pelos gestores dos órgãos licitantes, de modo a identificar mais rapidamente eventuais desconformidades em seus processos de contratação.

Palavras-chave: LLM; conformidade legal; licitação; tecnologia da informação

Abstract

This paper evaluates the potential use of Large Language Models to enhance the efficiency of preventive audit work in public procurement. Using the GPT-4o model, the study assesses the compliance of preliminary technical studies (ETPs) for IT solution procurements with the regulatory requirements of Instruction SGD/ME No. 94/2022 and compares the responses provided by the LLM with the analyses conducted by government internal auditors. The results demonstrate that automated analysis performs well in identifying non-compliances and provides well-founded analyses, taking less than a minute to evaluate each ETP, making it a powerful tool for increasing the productivity and efficiency of audit work. The work also highlights that the AI-implemented tool tends to indicate partial compliance in cases to which auditors argue for full compliance, mainly due to the extent of the regulatory requirements and the difficulty of contextual interpretation of the Instruction and the Preliminary Technical Studies. The main limitations identified include (1) the lack of public data on regulatory compliance evaluations conducted by audit bodies; (2) the consequent need to use human evaluations conducted by auditors individually rather than institutionally; (3) the relatively low number of ETPs analyzed by auditors for comparison with the automated solution; (4) the use of a single LLM; and (5) the fact that only some of the most relevant excerpts of the analyzed documents - not their full content - were provided to the language model. The results of this work can benefit audit teams at the federal, state, and municipal levels and have potential applications for any type of audited object, and not only for Information Technology. Finally, the AI-based solution proposed in this study can also be used directly by the managers of the bidding agencies to identify any non-compliances in their procurement processes more quickly.

Keywords: LLM; legal compliance; public bidding; information technology

Lista de figuras

Figura 1 - Diagrama da arquitetura do sistema computacional implementado para o teste com LLM.....	26
--	----

Lista de quadros

Quadro 1 - Matrizes de confusão do cenário 4 para os 7 testes de auditoria	56
--	----

Lista de tabelas

Tabela 1 - Cenários de testes realizados	27
Tabela 2 - Distribuição das respostas dos auditores por classe e por teste.....	28
Tabela 3 - Distribuição das respostas dos auditores e de cada cenário do LLM.....	29
Tabela 4 - Consolidação das métricas de avaliação do LLM nos 4 cenários	31
Tabela 5 - Exemplo de leitura dos valores de F1-Score para a classe Não".....	32
Tabela 6 - Comparação dos maiores valores de métricas para cada cenário.....	33
Tabela 7 - Tempo de execução da análise automatizada dos ETPs.....	34
Tabela 8 - Checklist de testes de auditoria sobre ETPs baseado na IN SGD/ME nº 94/2022	46
Tabela 9 - Modelos de prompt utilizados.....	47
Tabela 10 - Perguntas utilizadas para testes de auditoria realizados	49
Tabela 11 - Conjunto de instruções INST_01 para os testes de auditoria realizados	50
Tabela 12 - Conjunto de instruções INST_02 para os testes de auditoria	51
Tabela 13 - Conjunto de instruções INST_03 para os testes de auditoria realizados	53

Lista de abreviaturas e siglas

API	<i>Application programming interface</i>
ETP	Estudo Técnico Preliminar
IN	Instrução normativa
LLM	<i>Large Language Model</i> (Grande Modelo de Linguagem)
RAG	<i>Retrieval-Augmented Generation</i> (Geração aumentada por recuperação)
TI	Tecnologia da Informação

Sumário

1.	Introdução	17
2.	Referencial Teórico	18
2.1.	Processamento de Linguagem Natural (PLN)	18
2.2.	Grandes Modelos de Linguagem (LLM).....	19
2.3.	<i>Retrieval Augmented Generation</i> (RAG).....	21
2.4.	<i>Embeddings</i>	22
2.5.	Métricas para tarefas de classificação	23
3.	Metodologia	24
3.1.	Seleção e Análise Prévia de Estudos Técnicos Preliminares	24
3.2.	Implementação do ambiente computacional	25
4.	Resultados	27
4.1.	Discussão dos Resultados	34
4.1.1.	Controle/Teste 1	35
4.1.2.	Controle/Teste 2	36
4.1.3.	Controle/Teste 3	36
4.1.4.	Controle/Teste 4	37
4.1.5.	Controle/Teste 5	38
4.1.6.	Controle/Teste 6	38
4.1.7.	Controle/Teste 7	39
4.1.8.	Aspectos gerais	39
5.	Conclusão	41
	Referências bibliográficas.....	42
	Apêndice 1 – Controles da IN SGD/ME nº 94/2022 sobre Estudos Técnicos Preliminares.....	45
	Apêndice 2 – Modelos de Prompt, Testes (perguntas) e Instruções	47
	Apêndice 3 – Matrizes de Confusão para o Cenário 4	56

1. Introdução

Ao versar sobre o controle em processos de contratação da administração pública, a Lei 14.133/2021 (Nova Lei de Licitações - NLLC) determina, em seu artigo 169, que as contratações públicas devem seguir práticas contínuas e permanentes de gestão de riscos e controle preventivo, incluindo o uso de recursos de tecnologia da informação. Além disso, elas devem estar subordinadas ao controle social e sujeitar-se a diferentes linhas de defesa: servidores e empregados públicos, agentes de licitação e autoridades que atuam na estrutura de governança do órgão ou entidade (primeira linha); unidades de assessoramento jurídico e de controle interno do próprio órgão ou entidade (segunda linha); órgão central de controle interno da Administração e pelo Tribunal de Contas (terceira linha) (BRASIL, 2021).

No âmbito federal, o Órgão Central de Controle Interno é a Controladoria-Geral da União (CGU), a qual define, em sua Orientação Prática sobre Serviços de Auditoria, que a auditoria preventiva¹ em aquisições “tem por objetivo avaliar preventivamente processos de aquisição”, apresentando natureza preventiva para “mitigar riscos que podem impactar os objetivos das futuras contratações” (BRASIL, 2000; CONTROLADORIA-GERAL DA UNIÃO, 2022).

Dados do Portal de Relatórios de Auditoria da CGU mostram que, de janeiro de 2023 a julho de 2024, o Órgão realizou 158 auditorias preventivas a partir da geração de alertas pela ferramenta de análise automatizada de editais do poder executivo federal (ALICE) (CONTROLADORIA-GERAL DA UNIÃO, 2024b).

Esse tipo de ação de controle exige bastante esforço e atenção dos auditores, uma vez que todo o trabalho deve ocorrer entre a publicação do edital e a abertura da sessão pública da licitação.

No caso das contratações de Tecnologia da Informação na Administração Pública Federal, as licitações ocorrem basicamente na modalidade pregão eletrônico, a qual tem, por padrão, oito dias úteis entre a publicação do edital e a realização do certame. Sendo assim, as equipes de auditoria têm, no máximo, sete dias úteis para analisar: (1) aspectos legais relacionados a contratações públicas de maneira geral, previstos na NLLC; (2) questões de conformidade com as normas de contratações específicas para TI (no caso da administração direta, autárquica e fundacional, previstas na Instrução Normativa SGD/ME nº 94/2022 e nas demais portarias da Secretaria de Governo Digital, Órgão Central do SISP) (BRASIL, 2022); além da (3) adequação dos aspectos técnicos do objeto licitado às necessidades e condições dos órgãos licitantes, os quais evoluem muito rapidamente, no ritmo da tecnologia. O tempo disponível para as auditorias preventivas acaba sendo, portanto, insuficiente para a análise completa de todo o processo licitatório.

As avaliações de conformidade realizadas no âmbito das auditorias preventivas consistem, de forma resumida, na comparação do conteúdo dos editais e seus anexos frente aos dispositivos (ou controles) estabelecidos nos normativos legais e infralegais e, para esse tipo de tarefa, o emprego de tecnologias como processamento de linguagem natural (PLN), grandes modelos de linguagem (LLMs) e geração

¹ O art. 169, III da Lei 14.133/2021 atribui aos Órgãos Centrais de Controle Interno e aos Tribunais de Contas a competência para realização do **controle preventivo** das contratações públicas. No entanto, cada órgão de controle pode definir e, especialmente, nomear esses procedimentos de controle de forma própria. Neste trabalho, utilizam-se os termos “avaliação preventiva” e “análise preventiva” como sinônimos que se referem à etapa do controle preventivo compreendido entre a divulgação dos editais das licitações e a etapa de apresentação de propostas e lances desses certames.

aumentada por recuperação (*Retrieval-Augmented Generation* – RAG) apresenta grande potencial para automatização de boa parte do processo de trabalho das equipes de auditoria (CASELO; NUNES, 2024; LEWIS *et al.*, 2021; OECD, 2024; RAM *et al.*, 2023).

Assim, diante do tempo exíguo para a realização de auditorias preventivas sobre editais de licitação, e do recente e expressivo avanço das técnicas de inteligência artificial para processamento de textos não estruturados, bem como a potencial adequação dessas técnicas para automatização das atividades de auditoria, o presente estudo busca responder à seguinte questão: é possível a utilização de grandes modelos de linguagem (LLMs) para aumentar a eficiência das análises de conformidade legal em auditorias preventivas de contratações públicas?

Para essa investigação, este trabalho utiliza um grande modelo de linguagem para realizar a análise automatizada de um conjunto de Estudos Técnicos Preliminares de contratações de soluções de TI, comparando as respostas do modelo com análises fornecidas por auditores internos governamentais para os mesmos documentos.

Para avaliação do desempenho da solução automatizada, são utilizadas métricas automatizadas de avaliação de tarefas de classificação, bem como avaliação humana e medidas de tempo de execução das análises.

Os resultados deste estudo podem beneficiar equipes de auditoria tanto no nível federal quanto nas esferas estadual e municipal, além de terem potencial de aplicação para qualquer tipo de objeto auditado. Por fim, uma vez constatados a aplicabilidade e os benefícios do uso da inteligência artificial, ela pode ser empregada diretamente pelos gestores dos órgãos licitantes de modo a identificar mais rapidamente eventuais desconformidades em seus processos de contratação.

2. Referencial Teórico

2.1. Processamento de Linguagem Natural (PLN)

O Processamento de Linguagem Natural (PLN) é um campo de pesquisa dedicado a investigar e propor métodos computacionais para lidar com a linguagem humana. (CASELO; NUNES, 2024, cap. 1). O objetivo do PLN é desenvolver algoritmos e modelos capazes de compreender, gerar e manipular a linguagem humana de forma precisa e natural. (PATWARDHAN; MARRONE; SANSONE, 2023). A área se divide em duas grandes subáreas: Interpretação de Linguagem Natural (NLU), que tem como foco a análise e interpretação da linguagem, segmentando e classificando os componentes linguísticos para tentar apreender os significados construídos pelo ser humano; e a Geração de Linguagem Natural (NLG), que visa a gerar linguagem natural. Exemplos desses dois grandes processos podem ser reconhecidos na interação humana com um chatbot. No momento que o assistente virtual processa a mensagem do usuário para determinar a ação a ser tomada, tem-se a NLU, ao passo em que a geração da resposta propriamente dita configura a NLG. (CASELO; NUNES, 2024, cap. 1).

O PLN é utilizado nas mais diversas áreas, sendo facilmente reconhecido em aplicações de tradução automática, correção ortográfica e gramatical, assistentes virtuais, chatbots, sumarização automática, sistemas de recomendação, auxílio à escrita, classificação textual, recuperação de informação e detecção de fake news. Essas aplicações se beneficiam de recursos linguísticos como léxicos, dicionários,

corpus anotados, listas de frequências de palavras, taxonomias, ontologias e modelos de linguagem estatísticos ou neurais. (CASELO; NUNES, 2024, cap. 1).

Os trabalhos com Processamento de Linguagem Natural exigem, além de recursos computacionais como memória e processamento, um conjunto de dados textuais robusto e diversificado, algoritmos de aprendizado de máquina, conhecimento linguístico e de uso da língua. Destaca-se, no entanto, que estratégias automáticas de PLN têm limitações e geralmente processam caracteres, não unidades linguísticas, com base em padrões de ocorrência e contexto. Embora os modelos neurais, por exemplo, sejam poderosos, eles não "entendem" a língua no sentido humano, apenas reproduzem e extrapolam padrões aprendidos em *corpus* de treinamento (CASELO; NUNES, 2024, cap. 1).

Os benefícios do PLN incluem a capacidade de automatizar tarefas, extrair informações de grandes volumes de dados e facilitar a comunicação entre humanos e computadores. Os desafios residem na complexidade da linguagem humana, na ambiguidade inerente à linguagem natural e na dificuldade de representar o conhecimento semântico de forma completa e robusta. A busca por formalismos de representação híbridos, que combinem elementos simbólicos, estatísticos e neurais, é essencial para superar essas dificuldades e permitir um processamento de linguagem mais preciso e eficaz (CASELO; NUNES, 2024, cap. 1).

2.2. Grandes Modelos de Linguagem (LLM)

Um Modelo de Linguagem é uma representação matemática e computacional da língua, construído para capturar os padrões e nuances da linguagem humana. O objetivo principal de um Modelo de Linguagem é atribuir uma probabilidade a uma sequência de palavras, permitindo a predição de palavras futuras ou ausentes em um texto. A modelagem de linguagem é um componente fundamental na construção de sistemas de Processamento de Linguagem Natural, contribuindo para diversas tarefas, como tradução automática, geração de texto, análise de sentimentos e resposta a perguntas (CASELO; NUNES, 2024, cap. 15; DONG *et al.*, 2022; ZHAO *et al.*, 2023).

Os Grandes Modelos de Linguagem (LLMs) se destacam por possuírem uma enorme quantidade de parâmetros (bilhões ou até trilhões) e serem treinados em conjuntos de dados massivos e diversos, abrangendo diferentes domínios e línguas (CASELO; NUNES, 2024, cap. 15; ZHAO *et al.*, 2023). Essa escala massiva confere aos LLMs a capacidade de realizar tarefas complexas e exibir habilidades emergentes, como o aprendizado em contexto (*in-context learning*) e a estratégia de cadeia de pensamento (*chain-of-thought*), que simulam o raciocínio humano. O aprendizado em contexto permite que os LLMs resolvam tarefas sem treinamento adicional, apenas com instruções em linguagem natural e exemplos. Já a estratégia de cadeia de pensamento possibilita que os LLMs apresentem os passos intermediários de seu raciocínio, tornando suas respostas mais transparentes e compreensíveis (CASELO; NUNES, 2024, cap. 15).

Exemplos de LLMs incluem o GPT-3, GPT-4, PaLM, BLOOM, LLaMA e Sabiá. O ChatGPT e o Gemini, agentes de conversação populares, são exemplos de aplicações que utilizam LLMs para interagir com humanos de forma natural e realizar tarefas como tradução, geração de código, escrita criativa e resposta a perguntas complexas. Aplicações de LLMs não ficam restritas à área da computação, podendo ser encontradas nas mais diversas áreas de conhecimento, como Direito, Saúde

Mental e Educação (CASELO; NUNES, 2024, cap. 15; MINAEE *et al.*, 2024; SCHULHOFF *et al.*, 2024; ZHAO *et al.*, 2023).

Os LLMs podem ser classificados em modelos de código aberto (*open source*) ou de código fechado, de acordo com a disponibilidade do código-fonte e dos pesos desses modelos. LLMs *open source*, como o LLaMA e seus derivados (Alpaca, Vicuna), permitem que pesquisadores e desenvolvedores acessem, modifiquem e distribuam o código e os pesos do modelo. Essa abertura facilita a pesquisa, colaboração e adaptação dos modelos a diferentes tarefas e idiomas. Por outro lado, LLMs de código fechado, como o GPT-3, GPT-4 e o PaLM, são controlados por empresas e seus detalhes de implementação não são divulgados. O acesso a esses modelos se dá por meio de APIs, limitando a capacidade de pesquisa e personalização (MINAEE *et al.*, 2024; ZHAO *et al.*, 2023).

Para o desenvolvimento ou aperfeiçoamento de LLMs, são necessários recursos computacionais significativos, como grande capacidade de processamento (normalmente utilizando-se Unidades de Processamento Gráfico – GPUs) e memória, especialmente durante o treinamento, etapa durante a qual também deve estar disponível uma grande quantidade de dados de alta qualidade para que o trabalho seja concluído com sucesso. O conhecimento em PLN e aprendizado de máquina, bem como a familiaridade com frameworks específicos como TensorFlow e PyTorch, são essenciais para o desenvolvimento e ajuste fino (*fine tuning*) dos modelos. Por fim, plataformas como o Hugging Face facilitam o acesso e a utilização de LLMs pré-treinados (CASELO; NUNES, 2024, cap. 15; ZHAO *et al.*, 2023).

As vantagens de se trabalhar com LLMs incluem a capacidade de resolver tarefas complexas com alta precisão, a generalização para diferentes domínios e idiomas, a facilidade de uso por meio de prompts e a personalização por meio do *fine tuning*. No entanto, a utilização de grandes modelos de linguagem apresenta uma série de riscos, que devem ser tratados com bastante atenção.

Um dos principais problemas dos LLMs reside na sua propensão a perpetuar vieses presentes nos dados de treinamento. (OECD, 2024; SARKER, 2024; SOBRINO-GARCÍA, 2021). Se os dados refletem desigualdades sociais, o modelo pode replicar e amplificar essas desigualdades em suas saídas, resultando em decisões discriminatórias. (SOBRINO-GARCÍA, 2021; TOLEDO; MENDONÇA, 2023). A falta de transparência nos dados utilizados para treinar os LLMs comerciais agrava esse problema, dificultando a identificação e mitigação de vieses (OECD, 2024; WANG *et al.*, 2023).

Os LLMs podem também gerar informações falsas ou "alucinações", ou seja, conteúdo plausível, mas sem base factual (MUHLGAY *et al.*, 2023; SARKER, 2024; ZHAO *et al.*, 2023). Essas alucinações podem ser extremamente convincentes, tornando difícil para os usuários discernir a verdade. Por fim, a proliferação de informações falsas pode ter consequências negativas, minando a confiança nas informações e dificultando a tomada de decisões informadas. (OECD, 2024; ZHAO *et al.*, 2023).

A complexidade dos LLMs, especialmente por terem arquiteturas baseadas em redes neurais, torna difícil entender como eles chegam a determinadas conclusões. Essa falta de transparência dificulta a identificação de erros e vieses, dificultando a responsabilização por decisões equivocadas tomadas pelos modelos ou a partir de respostas deles (OECD, 2024; SARKER, 2024; TOLEDO; MENDONÇA, 2023; YOUNG *et al.*, 2021; ZHAO *et al.*, 2023).

Por fim, outro risco comumente associado ao uso de LLM se refere a privacidade de dados e segurança da informação. Os LLMs podem ser vulneráveis a

ataques de segurança e vazamento de informações confidenciais (OPENAI *et al.*, 2023; SARKER, 2024; YOUNG *et al.*, 2021). Dados sensíveis utilizados no treinamento podem ser extraídos por meio de técnicas de *prompt engineering*, colocando em risco a privacidade dos indivíduos (WANG *et al.*, 2023). A crescente integração dos LLMs em sistemas críticos exige medidas robustas de segurança para garantir a proteção de dados e a confiabilidade das aplicações.

Diante disso, faz-se necessário um uso responsável dos grandes modelos de linguagem, sempre conjugando a automatização e velocidade proferidas por eles com a supervisão e reflexão humana sobre os resultados obtidos. Além disso, a busca por garantia de ética, responsabilidade e transparência no desenvolvimento e uso de LLMs é fundamental para mitigar os riscos e garantir que esses modelos sejam utilizados para o bem da sociedade (CASELO; NUNES, 2024, cap. 15; SARKER, 2024).

2.3. Retrieval Augmented Generation (RAG)

Os Grandes Modelos de Linguagem representam um avanço significativo no campo do Processamento de Linguagem Natural. Entretanto, esses modelos enfrentam limitações quando se trata de acessar e integrar informações externas em suas operações, o que pode comprometer sua precisão e factualidade em cenários específicos. Neste cenário, a Geração Aumentada por Recuperação (RAG) surge como uma solução para esse desafio, combinando o poder generativo dos LLMs com a busca direcionada em bases de conhecimento externas (CASELO; NUNES, 2024, cap. 15; RAM *et al.*, 2023).

O RAG atua como uma ponte entre o conhecimento pré-treinado do LLM e a vastidão de informações disponíveis em fontes externas. Ao integrar a busca direcionada à geração de texto, o RAG permite que os LLMs acessem informações contextuais relevantes em tempo real, enriquecendo suas respostas e tornando-as mais precisas e informativas. Essa capacidade de integrar informações do mundo real torna os LLMs mais robustos e confiáveis, expandindo suas aplicações em áreas como resposta a perguntas, verificação de fatos e escrita criativa, em que a precisão factual é crucial (CASELO; NUNES, 2024, cap. 15; RAM *et al.*, 2023).

O funcionamento do RAG se baseia na interação entre um LLM pré-treinado e um sistema de recuperação de informação. Inicialmente, o LLM recebe uma entrada do usuário, que serve como base para a geração de texto. Em seguida, o sistema de busca é acionado para encontrar os documentos mais relevantes do *corpus*, considerando o contexto da tarefa e a entrada do usuário. A informação recuperada é então integrada ao LLM, permitindo que o modelo a utilize durante a geração da resposta (LEWIS *et al.*, 2021; RAM *et al.*, 2023).

Essa integração pode ser realizada de diferentes maneiras, dependendo da arquitetura específica do sistema RAG. Uma abordagem comum é incorporar os documentos recuperados como parte do contexto de entrada do LLM, permitindo que o modelo os processe juntamente com a entrada do usuário. Outra alternativa é utilizar a informação recuperada para guiar o processo de decodificação do LLM, influenciando a escolha das palavras e a estrutura da resposta (IZACARD *et al.*, 2022; LEWIS *et al.*, 2021).

As aplicações do RAG são diversas e demonstram seu potencial para revolucionar a forma como interagimos com a informação. Em tarefas de perguntas e respostas, o RAG permite que os sistemas forneçam respostas mais completas e precisas, acessando informações relevantes de um *corpus* de documentos. Em

geração de texto, o RAG pode ser utilizado para aprimorar a factualidade e a confiabilidade dos textos gerados, combatendo o problema da "alucinação" de informações (LEWIS *et al.*, 2021; MUHLGAY *et al.*, 2023; RAM *et al.*, 2023).

Apesar de suas vantagens, o trabalho com RAG também apresenta desafios. A complexidade do sistema, que envolve a busca, recuperação e integração de informações, exige expertise em diferentes áreas, como processamento de linguagem natural, recuperação de informação e aprendizado de máquina. O custo computacional da busca e recuperação de informações pode ser um obstáculo, especialmente se o *corpus* de documentos for extenso (RAM *et al.*, 2023).

Outro desafio importante é o risco de vieses presentes nos documentos utilizados como fonte de informação. É fundamental que os sistemas RAG sejam desenvolvidos com mecanismos para detectar e mitigar esses vieses, garantindo que as respostas geradas sejam justas e imparciais (FENG *et al.*, 2021; RAM *et al.*, 2023).

Em suma, o RAG representa um avanço promissor na área de geração de texto, com potencial para transformar a maneira como interagimos com a informação. Ao combinar a capacidade de linguagem dos LLMs com a riqueza de informações disponíveis em fontes externas, o RAG abre caminho para sistemas mais inteligentes, precisos e confiáveis. No entanto, é fundamental ter em mente os desafios e garantir que o desenvolvimento e a implementação do RAG sejam realizados de forma ética e responsável, priorizando a precisão, a transparência e a justiça.

2.4. *Embeddings*

Os *embeddings* são formas de representação de palavras e expressões como vetores. Essa técnica emergiu como uma alternativa poderosa pois, a partir da representação no espaço vetorial, foi possível desenvolver algoritmos para cálculo da similaridade de significado entre diferentes conjuntos de texto a partir de métricas de proximidade entre vetores (CHOWDHARY, 2020, cap. 18). Essa representação vetorial densa, aprendida a partir de grandes *corpora* textuais, permite que algoritmos capturem relações complexas entre as palavras, como sinonímia, antonímia e analogias (CASELO; NUNES, 2024).

Um modelo de *embedding* é um algoritmo que aprende a gerar esses vetores a partir de dados textuais. Diversos modelos de *embedding*, como Word2Vec (CHURCH, 2017), FastText (BOJANOWSKI *et al.*, 2017; FACEBOOK, 2022) e GloVe (PENNINGTON; SOCHER; MANNING, 2014) têm sido propostos, cada um com suas particularidades na forma como aprendem as representações vetoriais. O Word2Vec, por exemplo, analisa o contexto local das palavras. O FastText, por sua vez, se destaca por levar em conta a morfologia das palavras, representando-as como sequências de caracteres, o que permite gerar *embeddings* para palavras fora do vocabulário (CASELO; NUNES, 2024).

A utilização de *embeddings* revolucionou o campo do PLN, impactando positivamente o desempenho de diversas tarefas, como tradução automática, análise de sentimentos, sumarização e resposta a perguntas. Em tradução automática, por exemplo, *embeddings* bilíngues são utilizados para mapear palavras de diferentes idiomas em um espaço vetorial compartilhado, facilitando o processo de tradução. Na análise de sentimentos, *embeddings* auxiliam na identificação da polaridade das palavras, permitindo classificar textos como positivos, negativos ou neutros. As informações sobre aplicações de *embeddings* em análise de sentimentos não estão presentes nas fontes, mas foram inferidas a partir do conhecimento geral sobre PLN. A capacidade dos *embeddings* de capturar relações semânticas complexas e de

generalizar para palavras fora do vocabulário os torna uma ferramenta fundamental para o desenvolvimento de sistemas de PLN mais robustos e eficientes.

2.5. Métricas para tarefas de classificação

Em Processamento de Linguagem Natural (PLN), as tarefas de classificação consistem em atribuir uma categoria ou rótulo específico a um determinado dado de entrada. Esse dado pode ser um texto, um documento, uma imagem, entre outros.

Em que pese o fato de Grandes Modelos de Linguagem não serem dedicados especificamente a tarefas de classificação, sua capacidade de interpretação e sintetização de textos e a recente evolução de desempenho e eficácia desses LLMs os tornam bastante relevantes em tarefas de classificação da informação textual, senão de forma autônoma, pelo menos em conjunto a outras técnicas de inteligência artificial (PAIVA *et al.*, 2024; TOLEDO; MENDONÇA, 2023; ZHANG *et al.*, 2024; ZHAO *et al.*, 2023).

Entre as diferentes aplicações da classificação de texto, pode-se destacar a identificação da polaridade, quando se utiliza um conjunto limitado de rótulos; a anotação morfosintática, quando se utiliza um conjunto grande de categorias (mais de dez) para classificar as palavras de um texto de acordo com suas classes gramaticais e funções sintáticas; e a identificação de relações semânticas entre frases (CASELO; NUNES, 2024).

A avaliação da qualidade das tarefas de classificação é fundamental para garantir a eficácia dos sistemas de PLN. Essa avaliação se baseia em métricas que quantificam o desempenho do sistema e, entre as principais medidas de desempenho, destacam-se a acurácia, a precisão, o *recall* (abrangência ou cobertura) e o *F1-Score* (ou medida F).

A acurácia mede a proporção de classificações corretas realizadas por um sistema em relação ao total de casos analisados, ou seja, ela indica a porcentagem de acertos do sistema. A precisão avalia a proporção de classificações positivas corretas em relação ao total de classificações positivas realizadas. Já o *recall* mede a proporção de itens relevantes (positivos) que foram corretamente classificados. Por sua vez, o *F1-Score* combina precisão e o *recall* em uma única métrica, calculando a média harmônica entre as duas. É útil para equilibrar a importância da precisão e da abrangência na avaliação do sistema. (CASELO; NUNES, 2024).

Aliada às métricas acima, uma ferramenta importante de análise de desempenho de um sistema de classificação é a matriz de confusão. Trata-se de um recurso visual, organizado na forma de uma tabela, na qual as linhas representam as classes reais dos dados e as colunas representam as classes preditas pelo sistema. Cada célula da tabela indica a quantidade de dados que pertencem a uma determinada classe real e foram classificados em uma determinada classe predita (HUGGING FACE, 2024; SCIKIT LEARN, 2024). A matriz permite, assim: (1) identificar padrões de erro específicos do sistema, revelando quais classes são mais frequentemente confundidas entre si; (2) analisar problemas de classes desbalanceadas, mostrando o desempenho do sistema em cada classe individualmente; (3) comparar diferentes sistemas de classificação, permitindo a visualização das diferenças em seus padrões de acertos e erros.

Por fim, é importante destacar que a avaliação de um sistema de PLN que realiza classificação de texto exige uma abordagem cuidadosa e multifacetada. A escolha e a interpretação das métricas e ferramentas de avaliação devem ser guiadas pelos objetivos da avaliação, pelas características da tarefa e pelas propriedades dos

dados. Uma avaliação completa deve combinar métricas quantitativas com análises qualitativas, considerando tanto os acertos quanto os erros do sistema. É fundamental buscar um equilíbrio entre a precisão e o *recall* da avaliação, utilizando uma combinação estratégica de métricas e ferramentas que capturem os diferentes aspectos do desempenho do sistema (CASELO; NUNES, 2024).

3. Metodologia

O presente trabalho buscou avaliar a aplicabilidade de grandes modelos de linguagem na automatização de tarefas de auditoria interna governamental. Mais especificamente, buscou-se avaliar o desempenho de um LLM na realização automatizada de análise de conformidade de Estudos Técnicos Preliminares produzidos por diversas organizações públicas brasileiras em processos de planejamento de contratações de soluções de Tecnologia da Informação e Comunicação frente aos requisitos normativos definidos pela Instrução Normativa SGD/ME nº 94/2022.

Optou-se, neste estudo, pela avaliação de ETPs devido à relevância desse documento nos processos de contratações públicas. Nos termos do artigo 6º, XX da Lei 14.133/2021, trata-se de “documento constitutivo da primeira etapa do planejamento de uma contratação que caracteriza o interesse público envolvido e a sua melhor solução e dá base ao anteprojeto, ao termo de referência ou ao projeto básico a serem elaborados caso se conclua pela viabilidade da contratação” (BRASIL, 2021). Além disso, no caso específico das contratações de soluções de tecnologia da informação, a IN apresenta, em sua subseção II, de forma clara, requisitos específicos que devem constar dos Estudos Técnicos Preliminares (BRASIL, 2022).

Esses requisitos normativos foram, então, convertidos em um checklist de 7 (sete) testes de auditoria a serem aplicados sobre os ETPs. O objetivo da análise de conformidade é produzir, para cada um desses testes, uma resposta dividida em 2 partes: a inicial, na qual se indica, se o documento atende plenamente aos requisitos normativos, podendo a resposta ser “Sim”, “Não”, “Parcialmente” ou “Não se Aplica”; e a parte final, em que deve ser apresentada a fundamentação da parte inicial da resposta. O Apêndice 1 apresenta o checklist de auditoria elaborado para essa tarefa.

3.1. Seleção e Análise Prévia de Estudos Técnicos Preliminares

A fim de se avaliar o desempenho do LLM, é necessário comparar as respostas do modelo com respostas que seriam corretas (referências) para as avaliações.

À época de elaboração deste trabalho, não foram encontrados, nos Portais da Controladoria-Geral da União e do Tribunal de Contas da União (órgãos que auditam contratações públicas no Poder Executivo Federal), relatórios de auditorias que apresentassem avaliações de conformidade em contratações de TI.

Dessa forma, com auxílio da ferramenta Alice, da CGU, que analisa diariamente, de forma automatizada, os processos de compras e contratações públicas (CONTROLADORIA-GERAL DA UNIÃO, 2024a), foram selecionados todos os editais que a ferramenta classificou como sendo relacionados a TI entre janeiro e agosto de 2024, que totalizavam 263 registros. Foram efetivamente baixados e analisados os documentos de 115 processos, buscando especificamente os ETPs dessas contratações. Identificou-se, no entanto, que 59 dos casos não se adequavam

à análise pretendida, por: não se referirem, de fato, a objetos relativos a tecnologia da informação (25 casos); representarem, na verdade, ocorrências repetidas de processos que foram republicados (16 casos); não constituírem, ou apresentarem, de fato, o Estudo Técnico Preliminar (8 casos); apresentarem o conteúdo escaneado dos ETPs (7 casos), o que dificultaria a extração e processamento do conteúdo desses artefatos; apresentarem erro ao abrir o arquivo (2 casos); e conter apenas um anexo ao ETP (1 caso).

Assim, dos 56 editais restantes, 30 foram submetidos à análise de 2 auditores da Controladoria-Geral da União, os quais avaliaram os Estudos Técnicos Preliminares desses processos de contratação seguindo e preenchendo o checklist de auditoria da Tabela 8 do Apêndice 1.

As respostas dos auditores foram, então, confrontadas com as respostas produzidas pelo grande modelo de linguagem a partir da solução computacional implementada conforme descrito na seção a seguir.

3.2. Implementação do ambiente computacional

O ambiente computacional para teste de LLM na automatização de avaliações de conformidade legal foi implementado na plataforma Google Colab, por meio de programação em linguagem python.

Optou-se por utilizar uma arquitetura com RAG para que fossem repassados ao Modelo de Linguagem, além das instruções e testes a serem realizados, apenas os trechos mais relevantes, tanto do artefato analisado quanto do normativo de referência, já que tanto o normativo quanto os ETPs são documento extensos, normalmente de dezenas de páginas, e o envio completo deles poderia onerar bastante o sistema, principalmente pela necessidade de limitação da quantidade de tokens a serem enviados ao LLM e de um algoritmo para consolidação das respostas a partir do envio desses trechos de documentos em blocos separados.

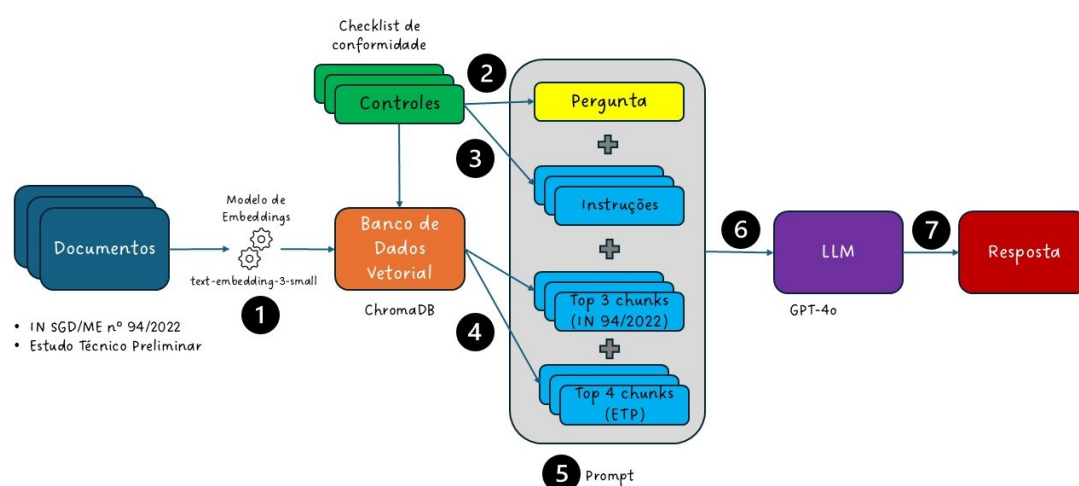
Além disso, as técnicas de RAG também se mostram efetivas em tarefas de perguntas e respostas para cenários em que há poucos exemplos a serem repassados ao modelo (IZACARD *et al.*, 2022), como é o caso deste trabalho, e viabilizam a utilização de modelos pré-treinados generalistas sem a necessidade de um novo treinamento (*fine tuning*) específico para o domínio da tarefa que se pretende realizar (RAM *et al.*, 2023; ZHAO *et al.*, 2023).

A Figura 1 apresenta o diagrama do ambiente computacional implementado, sendo os itens numerados nela as principais etapas ou configurações realizadas, conforme detalhado a seguir:

1. Inicialmente, o texto da Instrução Normativa SGD/ME nº 94/2022 foi dividido em trechos (chunks) correspondentes aos artigos no normativo. Ou seja, cada artigo da IN passou a representar um *chunk*. Esses trechos foram, então, convertidos em *embeddings* vetoriais utilizando-se o modelo text-embedding-ada-002 da OpenAI e, na sequência, armazenados e indexados em um banco de dados vetorial ChromaDB. O mesmo processo foi realizado para cada Estudo Técnico Preliminar analisado, porém, neste caso, os textos foram divididos em blocos de mesmo tamanho por meio da classe RecursiveCharacterTextSplitter da biblioteca LangChain.
2. Ainda numa etapa preparatória, os testes de auditoria, previstos no checklist da Tabela 8, do Apêndice 1, foram convertidos para o formato de perguntas a serem respondidas pelo Modelo de Linguagem.

3. Foram, então, definidas instruções a serem consideradas pelo LLM para a avaliação de cada controle normativo. Essas instruções são orientações sobre o que o Modelo de Linguagem deve levar em consideração na análise específica de cada controle.
4. A partir da pergunta que representa cada teste de auditoria, foram extraídos, por meio de uma busca por similaridade, os 3 principais trechos da IN e os 4 principais trechos de cada ETP relacionados ao controle que seria avaliado.
5. A pergunta, as instruções de avaliação e os trechos mais significativos do normativo e do artefato avaliado foram incluídos em um modelo de prompt.
6. Por meio da API da OpenAI, proprietária do GPT-4o, o prompt foi enviado, então, ao modelo.
7. A avaliação gerada pelo modelo foi recebida e armazenada para fins de comparação com a avaliação feita pelos auditores (respostas esperadas).

Figura 1 - Diagrama da arquitetura do sistema computacional implementado para o teste com LLM.



Fonte: elaborado pelo autor

A escolha do GPT-4o como LLM a ser utilizado se deve ao fato de ele ser reconhecidamente um representante do “estado da arte” entre os grandes modelos de linguagem, com bastante popularidade (CHIANG, 2024) e com uma boa relação entre desempenho (em termos de quantidade limite de token e velocidade na geração dos tokens) e custo por milhão de tokens gerados (ARTIFICIAL ANALYSIS, 2024). Além disso, trata-se de um modelo generalista e acessível pela Internet por meio de API por qualquer usuário ou organização, não requerendo infraestrutura computacional desses usuários para sua execução, e que oferece um painel de gestão de custos para acompanhamento financeiro da utilização do modelo.

O modelo de prompt foi estruturado de modo a orientar o LLM sobre o papel (persona) que ele deveria assumir para a tarefa, o estilo da resposta, a formatação estrutural dessa resposta, adequando-a à estrutura do checklist de auditoria, tendo como diretriz a pergunta referente ao teste de auditoria e, como informações adicionais, os trechos mais relevantes da IN SGD/ME nº 94/2022 e do ETP auditado. Assim, foram incluídos todos os elementos importantes de um bom prompt previstos por SCHULHOFF *et al.* (2024)

Uma vez que a indicação de conformidade legal ou não por meio das respostas “Sim”, “Não”, “Parcialmente” e “N/a” (parte inicial das respostas) pode ser considerada uma tarefa de classificação no âmbito do Processamento de Linguagem Natural, para fins de avaliação de desempenho do LLM, foram geradas matrizes de confusão e utilizadas métricas automáticas de acurácia, precisão, *recall* e *f1-score* para avaliação das respostas do modelo, tendo como referência as respostas dos auditores.

Além disso, a parte de explicação das respostas foi submetida a avaliação humana, por parte do autor, para confirmação sobre a correta estruturação delas, bem como análise das principais divergências encontradas entre as respostas produzidas pelo modelo de linguagem e as respostas esperadas.

Além das métricas automatizadas para tarefas de classificação, foi coletada, também, a informação do tempo de execução da análise automatizada dos ETPs.

Por fim, todo esse processo foi repetido em quatro cenários distintos, variando-se entre eles o modelo de prompt, o texto das perguntas e o detalhamento das instruções para cada teste, conforme resumido na Tabela 1.

Tabela 1 - Cenários de testes realizados

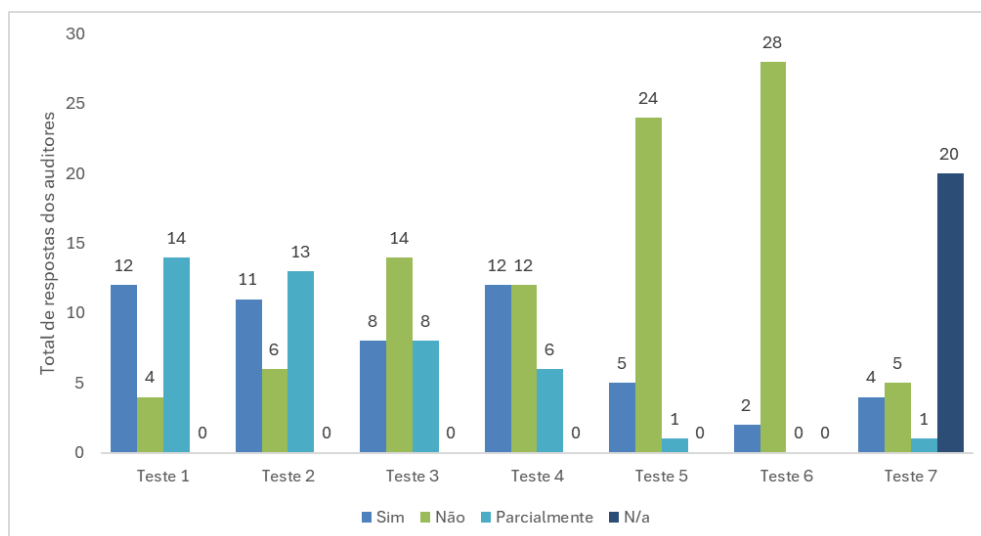
Cenário	Prompt	Testes	Instruções
1	PPT_01	PGT_01	Não fornecidas
2	PPT_02	PGT_02	INST_01
3	PPT_02	PGT_02	INST_02
4	PPT_02	PGT_02	INST_03

Fonte: elaborado pelo autor

Essas variações de configuração ocorreram a partir da análise humana do autor sobre as explicações das respostas do modelo e das matrizes de confusão geradas para cada cenário, e foram aplicadas no sentido de corrigir eventuais interpretações inadequadas do LLM e diminuir os erros cometidos por ele na análise de conformidade legal dos ETPs. O Apêndice 2 traz o detalhamento dos modelos de prompt, testes e instruções utilizados.

4. Resultados

A distribuição da parte inicial das respostas dos auditores sobre os 7 controles previstos na Instrução Normativa SGD/ME nº 94/2022 para os 30 Estudos Técnicos Preliminares analisados consta do Gráfico 1.

Gráfico 1 - Total e classe das respostas dos auditores por teste

Fonte: elaborado pelo autor

Observa-se que, para os testes 1 e 2, houve predominância de respostas “Sim” e “Parcialmente”, significando atendimento total ou parcial aos requisitos da IN. Já nos testes 3 e, especialmente, 5 e 6, houve predominância de respostas “Não”, indicando a não conformidade normativa. O teste 4 teve resultados iguais para “Sim” e “Não” e menor quantidade de “Parcialmente”. Por fim, apenas para o teste 7 observou-se a ocorrência de “N/a” (Não se aplica), classe que representou a maioria absoluta das respostas dos auditores. Dessa forma, a distribuição das respostas de referência ficou como apresentado na Tabela 2:

Tabela 2 - Distribuição das respostas dos auditores por classe e por teste

Teste / Classe	Sim	Não	Parcialmente	N/a
Teste 1	40,0%	13,3%	46,7%	0,0%
Teste 2	36,7%	20,0%	43,3%	0,0%
Teste 3	26,7%	46,7%	26,7%	0,0%
Teste 4	40,0%	40,0%	20,0%	0,0%
Teste 5	16,7%	80,0%	3,3%	0,0%
Teste 6	6,7%	93,3%	0,0%	0,0%
Teste 7	13,3%	16,7%	3,3%	66,7%
TOTAL	25,7%	44,3%	20,5%	9,5%

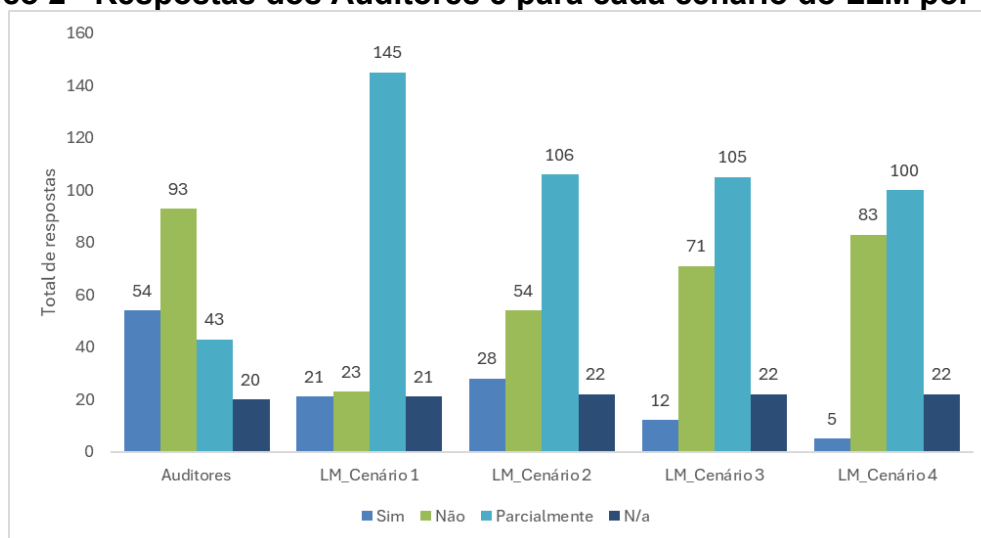
Fonte: elaborado pelo autor

Verifica-se que, no total, 25,7% das respostas foram “Sim”, sendo os testes 1 e 4 aqueles que tiveram a maior proporção dessa classe para os ETPs analisados (40%). A classe “Não” teve a maior proporção de respostas no total, com 44,3%, chegando a representar 93,3% das respostas no caso do teste 6. A resposta

“Parcialmente” foi dada em 20,5% das análises dos auditores, sendo o teste 1 aquele em que essa categoria foi usada na maior proporção (46,7%). Por fim, apenas no teste 7 foi utilizada a opção “N/a”, que representou, para esse teste específico, 66,7% das respostas e, no geral, apenas 9,5% das classes utilizadas pelos auditores.

Quanto às respostas fornecidas pelo LLM nos diferentes cenários, o Gráfico 2 apresenta o total de respostas por classe para cada cenário:

Gráfico 2 - Respostas dos Auditores e para cada cenário do LLM por classe



Fonte: elaborado pelo autor

É possível constatar que a mudança no modelo de prompt, nas perguntas e nas instruções promoveram mudanças significativas na análise automatizada. Respostas do tipo “Parcialmente” caíram de 145 para 100 ocorrências do cenário 1 para o cenário 4. Também houve redução no total de respostas “Sim” fornecidas pelo modelo, as quais somavam 21 ocorrências no cenário 1, chegaram a 28 no cenário 2, mas foram diminuindo até chegar em apenas 5 ocorrências no cenário 4. Por sua vez, a classe “Não”, que, no cenário 1, registrou apenas 23 respostas, teve sua frequência progressivamente ampliada, até chegar em 83 ocorrências no cenário 4. Por fim, houve estabilidade para respostas do tipo “N/a”, que começaram com 21 ocorrências no cenário 1 e, em todos os demais, ocorreram 22 vezes.

A distribuição percentual do total de respostas dos auditores e do grande modelo de linguagem está indicada na Tabela 3:

Tabela 3 - Distribuição das respostas dos auditores e de cada cenário do LLM

	Auditores	LM_Cenário 1	LM_Cenário 2	LM_Cenário 3	LM_Cenário 4
Sim	25,7%	10,0%	13,3%	5,7%	2,4%
Não	44,3%	11,0%	25,7%	33,8%	39,5%
Parcialmente	20,5%	69,0%	50,5%	50,0%	47,6%
N/a	9,5%	10,0%	10,5%	10,5%	10,5%

Fonte: elaborado pelo autor

Observa-se que, enquanto no caso dos auditores, a maior proporção de respostas foi “Não” (44,3%), no caso do LLM, a classe “Parcialmente” foi sempre a preponderante, mesmo no cenário 4, em que essa proporção foi a menor, ocorrendo em 47,6% das análises. Também é possível verificar que, em nenhum dos casos avaliados automaticamente, houve uma distribuição uniforme das respostas entre as classes (25% das respostas para cada uma das quatro classes).

No entanto, a distribuição geral das respostas por classe não indica, por si só, se o desempenho da análise automatizada está adequado. Para isso, é preciso avaliar o quanto dessas respostas estão corretas, bem como os tipos de erros que estão sendo cometidos. A Tabela 4 apresenta um resumo geral das métricas de avaliação automatizada, segmentadas por teste de auditoria, para os 4 cenários de uso do LLM.

Em azul escuro, foram destacados os valores máximos observados para cada métrica, podendo esses valores ocorrerem em mais de um cenário. Em azul claro, foram realçados os valores das métricas que, mesmo não sendo os valores máximos observados, foram os maiores valores observados para um teste específico.

Tabela 4 - Consolidação das métricas de avaliação do LLM nos 4 cenários

Cenário	Teste	Acurácia Simples	Precisão Sim	Precisão Não	Precisão Parcialmente	Precisão N/a	Recall Sim	Recall Não	Recall Parcialmente	Recall N/a	F1-Score Sim	F1-Score Não	F1-Score Parcialmente	F1-Score N/a
1	1	0,467	0,000	0,000	0,467		0,000	0,000	1,000		0,000	0,000	0,636	
	2	0,433	0,500	0,000	0,400		0,455	0,000	0,615		0,476	0,000	0,485	
	3	0,333	1,000	0,667	0,269		0,125	0,143	0,875		0,222	0,235	0,412	
	4	0,300	0,000	0,750	0,231		0,000	0,250	1,000		0,000	0,375	0,375	
	5	0,533	0,500	0,938	0,000		0,200	0,625	0,000		0,286	0,750	0,000	
	6	0,000	0,000	0,000	0,000		0,000	0,000	0,000		0,000	0,000	0,000	
	7	0,700	0,250	0,000	0,000	0,905	0,500	0,000	0,000	0,950	0,333	0,000	0,000	0,927
2	1	0,467	0,500	0,000	0,462		0,167	0,000	0,857		0,250	0,000	0,600	
	2	0,367	0,353	0,000	0,385		0,545	0,000	0,385		0,429	0,000	0,385	
	3	0,367	0,000	0,600	0,250		0,000	0,429	0,625		0,000	0,500	0,357	
	4	0,300	1,000	0,750	0,200		0,083	0,250	0,833		0,154	0,375	0,323	
	5	0,567	0,333	1,000	0,100		0,400	0,583	1,000		0,364	0,737	0,182	
	6	0,733	0,000	0,957	0,000		0,000	0,786	0,000		0,000	0,863	0,000	
	7	0,733	0,000	1,000	0,000	0,864	0,000	0,600	0,000	0,950	0,000	0,750	0,000	0,905
3	1	0,467	0,500	0,000	0,455		0,333	0,000	0,714		0,400	0,000	0,556	
	2	0,400	0,000	0,000	0,414		0,000	0,000	0,923		0,000	0,000	0,571	
	3	0,233	0,000	0,400	0,200		0,000	0,143	0,625		0,000	0,211	0,303	
	4	0,533	1,000	0,786	0,267		0,083	0,917	0,667		0,154	0,846	0,381	
	5	0,667	0,000	0,833	0,000		0,000	0,833	0,000		0,000	0,833	0,000	
	6	0,833	0,000	1,000	0,000		0,000	0,893	0,000		0,000	0,943	0,000	
	7	0,700	0,000	0,667	0,000	0,864	0,000	0,400	0,000	0,950	0,000	0,500	0,000	0,905

Cenário	Teste	Acurácia Simples	Precisão Sim	Precisão Não	Precisão Parcialmente	Precisão N/a	Recall Sim	Recall Não	Recall Parcialmente	Recall N/a	F1-Score Sim	F1-Score Não	F1-Score Parcialmente	F1-Score N/a
4	1	0,467	0,750	0,000	0,458		0,250	0,000	0,786		0,375	0,000	0,579	
	2	0,467	0,000	1,000	0,448		0,000	0,167	1,000		0,000	0,286	0,619	
	3	0,433	0,000	0,667	0,278		0,000	0,571	0,625		0,000	0,615	0,385	
	4	0,500	0,000	0,786	0,250		0,000	0,917	0,667		0,000	0,846	0,364	
	5	0,667	0,000	0,833	0,000		0,000	0,833	0,000		0,000	0,833	0,000	
	6	0,933	1,000	0,964	0,000		0,500	0,964	0,000		0,667	0,964	0,000	
	7	0,667	0,000	0,500	0,000	0,864	0,000	0,200	0,000	0,950	0,000	0,286	0,000	0,905

Fonte: elaborado pelo autor

A Tabela 5 apresenta um exemplo de leitura da métrica F1-score para a classe “Não”:

Tabela 5 - Exemplo de leitura dos valores de F1-Score para a classe “Não”

Teste	Cenário em que foi observado o valor máximo	Valor de f1-score para “Não”
1	Não houve. Todos os valores foram 0	0,000
2	4	0,286
3	4	0,615
4	3 e 4	0,846
5	3 e 4	0,833
6	4	0,964
7	2	0,750

Neste caso da métrica de F1-Score para a classe “Não”, o valor máximo foi 0,964, observado no cenário 4 para o teste 6. Ainda para a mesma métrica, mas considerando os demais 6 testes, os maiores valores foram observados: para os testes 2 e 3, no

cenário 4 (0,286 e 0,615, respectivamente); para os testes 4 e 5, nos cenários 3 e 4 (0,846 e 0,833); e, para o teste 7, no cenário 2 (0,750). O teste 1 teve apenas valores iguais a 0 em todos os cenários.

A Tabela 4 permite, ainda, uma análise do desempenho geral do LLM ao longo dos cenários com base nas métricas automatizadas e os destaques feitos conforme os valores delas. A Tabela 6 apresenta essa análise:

Tabela 6 - Comparação dos maiores valores de métricas para cada cenário

Cenário	Total de métricas com o valor máximo para a classe	Total de Métricas com o maior valor para a classe para um determinado controle	% de Métricas com valor > 0,6 ²	% de Métricas com valor > 0,7 ²	% de Métricas com valor > 0,8 ²	% de Métricas com valor > 0,9 ²
1	7	12	22,73%	15,15%	12,12%	10,60%
2	6	10	28,79%	25,76%	16,67%	10,60%
3	3	12	33,33%	25,76%	21,21%	10,60%
4	8	19	40,91%	28,79%	24,24%	15,20%

Fonte: elaborado pelo autor:

Constata-se que o cenário 4 apresentou a maior quantidade de métricas em que se observou o valor máximo para toda a classe (8), bem como a maior quantidade de métricas cujo valor foi o maior considerando apenas um controle específico (19). Além disso, esse cenário também teve o maior percentual de métricas com valores acima de 0,6, 0,7, 0,8 e 0,9 entre todos os cenários implementados.

Quanto ao tempo de execução, a Tabela 7 apresenta a duração da análise automatizada dos 30 ETPs.

² Para fins de cálculo do percentual, foram excluídas do total de contagens as medidas de precisão, recall e f1-score para a classe “N/a” dos testes 1 a 6, já que essa classe só ocorreu para o teste 7. Considerando que o cálculo do percentual se destina a uma comparação relativa dos cenários, considera-se que essa exclusão não traz prejuízo à análise

Tabela 7 - Tempo de execução da análise automatizada dos ETPs

Cenário	Duração (min)
1	19,35
2	20,65
3	22,36
4	25,92

Fonte: elaborado pelo autor

Do cenário 1, mais simples, para o cenário 4, com modelos de prompt, perguntas e, sobretudo, instruções maiores, houve um aumento do tempo de execução de 19,35 para 25,92 minutos.

Considerando, então, o objetivo deste trabalho, qual seja, avaliar a possibilidade de utilização de grandes modelos de linguagem (LLMs) para aumentar a eficiência das análises de conformidade legal em auditorias preventivas de contratações públicas, o restante desta seção, bem como a seção 4.1 -

Discussão dos Resultados, considerarão os resultados apresentados no Cenário 4.

Conforme apresentado na Tabela 4, a acurácia da avaliação pelo LLM ficou entre 0,433 e 0,467 nos primeiros 3 testes, 0,5 no teste 4, e acima de 0,66 nos demais testes, chegando ao máximo de 0,933 no teste 6.

Utilizando-se a métrica de F1-Score para avaliação geral do desempenho de classificação do modelo de linguagem, já que ela representa uma média harmônica entre precisão e recall, os dados da tabela mostram que, para a classe “Sim”, o F1-Score ficou em 0,667 para o controle 6, 0,375 no teste 1 e 0 nos demais testes.

Para a classe “Não”, a ferramenta de Inteligência Artificial apresentou F1-Score nulo no teste 1, 0,286 para os controles 2 e 7. 0,615 no teste 3 e acima de 0,83 nos demais casos, chegando a 0,964 para o teste 6.

Quanto à indicação de conformidade parcial, o modelo apresentou F1-Score igual a 0 nos testes 6, 7 e 8, pouco acima de 0,36 nos controles 3 e 4, 0,579 no teste 1 e 0,619 para o controle 2.

Por fim, a classe “N/a” foi utilizada apenas para o teste 7, tendo sido atingido um F1-Score de 0,905.

As matrizes de confusão geradas para as análises do Cenário 4 são apresentadas no Apêndice 3.

4.1. Discussão dos Resultados

Considerando, inicialmente, os valores de acurácia, que representa a proporção de classificações corretas da ferramenta em relação ao total de classificações, e detalhando-a por cada testes podemos ter uma visão geral do desempenho da ferramenta.

Para os testes 1, 2 e 3, a ferramenta apresentou acurácia relativamente baixa, com valores em torno de 0,45. Isso indica que a ferramenta teve dificuldades em classificar corretamente os artefatos para esses controles específicos.

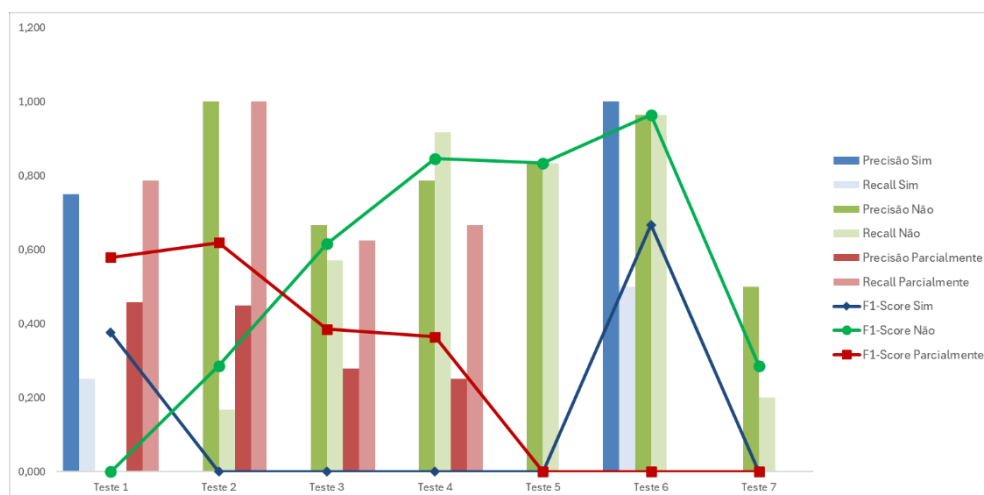
No teste 4, a acurácia de 0,5 indica que a ferramenta acertou a classificação em metade dos casos, o que pode ser considerado intermediário, com espaço para melhorias.

Já para os controles 5 e 7, a ferramenta apresentou uma acurácia de 0,667, o que representa um desempenho considerável.

Por fim, no teste 6, a ferramenta obteve a maior acurácia observada, 0,933, indicando um alto índice de classificações corretas.

No entanto, para se compreender melhor os testes e classes em que o LLM apresentou maior dificuldade ou assertividade na classificação, é necessário complementar a análise de acurácia com avaliação das métricas de precisão, recall e F1-Score. Gráfico 3 ajuda a visualizar o desempenho da ferramenta nesses itens.

Gráfico 3 - Precisão, Recall e F1-Score para classes "Sim", "Não" e "Parcialmente"



Fonte: elaborado pelo autor

4.1.1. Controle/Teste 1

O teste 1 avalia se "As necessidades de negócio e tecnológicas e os requisitos necessários e suficientes à escolha da solução de TIC foram definidos?". A ferramenta apresentou desempenho inconsistente para este controle. Apesar da precisão relativamente baixa (0,458), observou-se um alto recall (0,786) para a classe "Parcialmente", indicando que a ferramenta foi capaz de identificar a maioria dos artefatos com conformidade parcial.

A precisão de 0,750 para a classe "Sim" indica que, dos artefatos classificados como conformes pela ferramenta para este controle, 75% estavam realmente em conformidade total. No entanto, o baixo recall (0,250) revela que a ferramenta só identificou 25% dos artefatos realmente conformes para este controle. Isso resultou em um F1-score de 0,375, que representa um baixo equilíbrio entre precisão e recall para a constatação de conformidade.

Para a classe "Não", o desempenho foi deficiente, com precisão, recall e F1-score nulos. Isso sugere que a ferramenta tem dificuldades em reconhecer a conformidade total e a não conformidade para este controle, sendo mais eficaz na identificação de conformidades parciais.

A avaliação da matriz de confusão para esse controle, constante do Quadro 1, revela que o erro mais cometido pelo LLM (9 ocorrências) foi considerar ETPs apenas parcialmente atendidos quando os auditores os consideraram integralmente em conformidade.

A partir da avaliação humana desses erros mais frequentes, constatou-se que os motivos indicados pelo modelo de linguagem para a conformidade parcial, e não

total, concentraram-se em 3 fatores: ausência da definição das necessidades de negócio (citada em 3 casos), falta de clareza e detalhamento (4 ocorrências) e indicação de necessidades ou requisitos genéricos, e não específicos dos órgãos contratantes (citada 5 vezes).

Assim, observa-se que a ferramenta apresenta uma avaliação mais rigorosa do que os auditores em relação ao teste 1. Isso pode estar associado aos termos “necessários e suficientes” da pergunta feita para o teste, já que a determinação do que é se enquadra nesses conceitos é subjetivo, depende de cada solução contratada e do contexto do órgão contratante. Assim, os auditores podem utilizar, em seu julgamento sobre os ETPs, informações que não constam desses documentos.

4.1.2. Controle/Teste 2

O controle 2 verifica se “A solução tecnológica escolhida (objeto da contratação) resolve o problema do órgão ou entidade e/ou atende à necessidade descrita no Documento de Formalização da Demanda (DFD)”. Similarmente ao Controle/Teste 1, a ferramenta apresentou desempenho inconsistente, com um alto recall (1,000) para a classe “Parcialmente”, indicando a identificação de todos os artefatos com conformidade parcial.

Para a classe “Sim”, o desempenho foi ruim, com valores de precisão, recall e F1-score zerados, o que mostra que a ferramenta não conseguiu distinguir os casos de conformidade total para esse teste.

No caso das não conformidades, apesar da alta precisão (1,000), o recall foi baixo (0,167), mostrando que a ferramenta identificou apenas uma minoria dos casos de desconformidade.

Assim como no caso do controle 1, a matriz de confusão indica que, para o teste 2, o erro mais frequente cometido pelo LLM foi atribuir a classe “Parcialmente” aos casos em que os auditores indicaram a conformidade total (classe “Sim”), o que ocorreu em 11 casos. Porém, o teste 2 apresenta particularidades.

A primeira dificuldade da solução automatizada pode estar relacionada ao fato de que a pergunta do teste remete a um outro documento, que não o ETP, o DFD, e que não foi fornecido para a análise. Buscou-se contornar essa limitação indicando, nas instruções, que o modelo comparasse a solução escolhida frente às necessidades apresentadas no próprio ETP. Ainda assim, os resultados não foram satisfatórios.

A avaliação humana apontou que o motivo principal pelo qual a ferramenta reduziu o juízo de conformidade para parcial foi o fato de não ter ficado demonstrado de forma clara que a solução escolhida atende aos requisitos específicos dos contratantes (indicado em 10 dos 11 casos). Porém, percebeu-se que os auditores, na maioria dos casos em que afirmaram a conformidade integral, justificaram seu posicionamento com base em várias seções dos documentos analisados (3 ou 4 seções).

Isso demonstra uma dificuldade da solução automatizada em identificar a relação entre seções distintas do documento que possam formar, em conjunto, uma justificativa adequada para a indicação de conformidade integral com o requisito normativo. Como tentativas de se aprimorar o desempenho nesse teste, podem ser feitos novos aprimoramentos nas instruções fornecidas ao modelo de linguagem, ou mesmo a ampliação da quantidade de *chunks* do artefato retornados pela busca por similaridade a partir da pergunta do teste.

4.1.3. Controle/Teste 3

O controle 3 avalia se "Consta do Estudo Técnico Preliminar, de forma detalhada, motivada e justificada, inclusive quanto à forma de cálculo, o quantitativo de bens e serviços necessários para a composição da solução de TIC". Nessa análise, a ferramenta teve desempenho mediano para a classe "Não", com precisão de 0,667 (dois terços do que o modelo indicou estar em desconformidade realmente estavam) e recall de 0,571, indicando que a solução conseguiu identificar pouco mais que a metade dos casos de não conformidade.

Para a classe "Parcialmente", a ferramenta apresentou desempenho relativamente ruim, com recall de 0,625, mas precisão e F1-score mais baixos.

Para a classe "Sim", o desempenho foi nulo, similarmente ao controle 2, o que mostra que a ferramenta não conseguiu identificar os casos de conformidade total para esse teste.

Novamente, assim como nos 2 primeiros testes, a matriz de confusão mostra que, para o controle 3, a maior quantidade de erros detectados foi em casos para os quais os auditores indicaram a conformidade total, mas a ferramenta considerou o atendimento apenas parcial (7 ocorrências).

A avaliação humana desses erros constatou que, em todos eles, o LLM reduziu o grau de conformidade devido à ausência de uma memória de cálculo que mostrasse de forma clara como os quantitativos de itens ou serviços a serem contratados foram estimados. Tal interpretação, no entanto, parece correta, dado o comando da pergunta do teste, que exige uma forma "detalhada, motivada e justificada, inclusive quanto à forma de cálculo" para a explicação da quantidade de bens e serviços. Neste caso os auditores apresentaram opiniões mais flexíveis, normalmente indicando que as informações presentes nas seções sobre "Estimativa da Demanda" eram suficientes para o atendimento integral ao requisito normativo.

Destaca-se, no entanto, que, a indicação de conformidade parcial parece ser uma tendência da solução automatizada pois, mesmo em casos em que os auditores opinaram pela não conformidade, o modelo atribuiu a classe "Parcialmente" (erro ocorrido em 6 casos). Nessas situações, os auditores detectaram insuficiência das justificativas apresentadas, enquanto a ferramenta demonstrou opinião mais flexível pelo fato de haver uma "tentativa" de justificativa.

Assim, constata-se que houve uma dificuldade da solução automatizada em delimitar bem a desconformidade de situações de conformidade parcial. Melhorias nas instruções podem contribuir para o aprimoramento da análise para esse teste.

4.1.4. Controle/Teste 4

Este controle verifica se "No Estudo Técnico Preliminar, foi realizada análise comparativa de soluções, observando os aspectos econômicos e qualitativos em termos de benefício para o alcance dos objetivos da contratação". A ferramenta apresentou um bom desempenho na classe "Não", com precisão de 0,786, recall de 0,917 e F1-score de 0,846. Isso indica que a ferramenta foi eficaz na detecção de não conformidades para este controle.

O modelo não foi capaz de identificar nem um caso de conformidade total, apresentando com valores de precisão, recall e F1-score zerados para esse controle.

Para a classe "Parcialmente", a ferramenta apresentou um desempenho intermediário, com recall acima de 0,625, mas precisão e F1-score mais baixos.

Mais uma vez, a maior quantidade de erros ocorreu quando o modelo de linguagem atribuiu aos ETPs uma conformidade parcial ("Parcialmente"), enquanto os

auditores indicaram a conformidade total ("Sim"), o que aconteceu em 11 casos, conforme pode ser observado na matriz de confusão para esse teste.

Por meio da avaliação humana, percebeu-se os referidos erros ocorreram devido à extensão do inciso que serve de base para o teste 4 – inciso II do artigo 11 da Instrução Normativa SGD/ME nº 94/2022. Trata-se de um dispositivo com 10 alíneas, cada uma citando um aspecto que deve ser observado na análise comparativa de soluções. Assim, mesmo quando a ferramenta identificou nos ETPs alguns ou mesmo a maior parte desses itens, a não detecção de qualquer um deles já foi suficiente para que o modelo considerasse a conformidade apenas como parcial. Isso se refletiu na matriz de confusão, por exemplo, no fato que nenhum artefato foi considerado pelo LLM como estando em conformidade total.

Por outro lado, a análise dos auditores é mais flexível e, assim como no caso do teste 1, pode levar em conta elementos subjetivos ou o conhecimento do objeto, por exemplo, para formar o julgamento, de modo que eles podem identificar que os requisitos normativos de algumas das alíneas não se aplicam aos casos concretos em avaliação.

Trata-se, portanto, de uma dificuldade de interpretação do normativo frente aos casos concretos. No entanto, o fato de ter sido identificado pela ferramenta que os ETPs deixam de mencionar todos os aspectos cobrados pelo inciso I do artigo 11 da IN, mesmo que para indicar sua não aplicabilidade, pode servir de alerta para os órgãos responsáveis pela normatização, já que as exigências da IN podem estar muito extensas ou os órgãos contratantes podem não estar dispondo de capacidade suficiente para realização da análise solicitada pela norma.

4.1.5. Controle/Teste 5

Este controle avalia se "No Estudo Técnico Preliminar, foi realizada a análise comparativa de custos, por meio da comparação de custos totais de propriedade (Total Cost Ownership - TCO) das soluções consideradas viáveis". Nele, a ferramenta apresentou bom desempenho para a classe "Não", com precisão, recall e F1-score iguais a 0,833, indicando boa capacidade de identificação de não conformidades.

Já para as classes "Sim" e "Parcialmente", o desempenho foi nulo, revelando a dificuldade da ferramenta na identificação de conformidades totais e parciais para esse teste.

Cumprir destacar, no entanto, que houve apenas uma resposta dos auditores indicando a conformidade parcial, conforme apresentado no Gráfico 1. Assim, o desempenho da ferramenta para essa classe deve ser relativizado, e não deve ser considerado ruim.

A análise da matriz de confusão para o controle 5 mostra que não houve um tipo de erro específico com ocorrência destacadamente maior à dos demais, o que, aliado à acurácia considerável (0,667) e ao fato de que as respostas dos auditores foram majoritariamente pela classe "Não" nesse teste, mostra que o LLM teve um bom desempenho geral nessa avaliação.

4.1.6. Controle/Teste 6

O teste 6 verifica se "No Estudo Técnico Preliminar, a forma de pagamento (remuneração) escolhida encontra-se adequadamente fundamentada conforme critérios técnicos e financeiros". A ferramenta apresentou excelente acurácia (0,933) para este controle, com um desempenho notável na classe "Não", com valores de

precisão, recall e F1-score de 0,964, demonstrando grande confiabilidade do LLM na identificação de desconformidades.

Para a classe "Sim", houve desempenho satisfatório, com precisão de 1, recall de 0,5 e F1-score de 0,667, indicando capacidade razoável de identificar conformidades totais.

Já para a classe "parcialmente", o desempenho foi nulo, assim como no teste 5. Entretanto, no caso do teste 6, observa-se, pelo Gráfico 1, que esse desempenho deve ser considerado, na verdade, positivo, já que, de fato, não houve resposta dos auditores indicando conformidade parcial.

A matriz de confusão para o controle 6 revela que houve apenas 3 tipos de erros da ferramenta, cada um com apenas uma ocorrência, o que pode ser considerado excelente.

Também para esse teste, as respostas dos auditores indicaram não conformidade na grande maioria dos casos (28 de 30, conforme apresentado no Gráfico 1), o que pode ter levado a solução automatizada a apresentar, neste caso, seu melhor desempenho entre todos os controles e cenários avaliados.

4.1.7. Controle/Teste 7

O controle 7 avalia se "Caso o Estudo Técnico Preliminar (ETP) esteja propondo a contratação por meio da adesão a Sistema de Registro de Preços (SRP) ou Ata de Registro de Preços (ARP), consta desse ETP registro do ganho de eficiência, da viabilidade e da economicidade para a administração pública federal da utilização da ata ou do sistema de registro de preços". Trata-se de um teste que apresenta uma particularidade, já que foi o único em que havia a previsão de uso da classe "N/a", indicando que o controle não se aplicava. E, justamente para essa classe, a ferramenta apresentou seu melhor desempenho no teste, com precisão de 0,864, recall de 0,95 e F1-score igual a 0,905. Ou seja, o LLM apontou, com alta confiabilidade, os ETPs para os quais o controle 7 não se aplicava.

Para as classes "Sim" e "Parcialmente", a ferramenta apresentou desempenho nulo, não conseguindo identificar nenhum caso de conformidade total ou parcial.

Por fim, para a classe "Não", o desempenho foi limitado, com valores baixos de precisão e recall.

Assim como no caso do controle 5, a análise da matriz de confusão mostra que não houve um tipo de erro específico com ocorrência destacadamente maior à dos demais. Houve, no entanto, uma maior quantidade de tipos de erros diferentes – 6 no total – mas com ocorrências baixas (máximo de 3), o que pode ser considerado aceitável.

4.1.8. Aspectos gerais

A partir dos casos específicos de cada teste individual, pode-se observar que a ferramenta apresentou um ótimo desempenho na identificação de não conformidades, notadamente para os controles 3, 4, 5 e 6, sendo esse último aquele em que houve o melhor desempenho de todos.

O Gráfico 1 mostra que, justamente nestes controles, os auditores indicaram a classe "Não" de forma predominante em suas respostas, favorecendo, portanto, o bom desempenho geral do modelo nesses casos.

Já o Gráfico 2 mostra que, com as mudanças de prompt, textos das perguntas e das instruções para realização dos testes, a quantidade total de respostas "Não" do

LLM aumentou progressivamente, atingindo o máximo no cenário 4. A Tabela 4, por sua vez, permite constatar que, do cenário 1 para o cenário 4, o desempenho do modelo de linguagem para a classe “Não” foi melhorando significativamente, com o F1-score para a classe apresentando o maior valor de 5 dos 7 testes em todos os cenários. Assim, pode-se inferir que as mudanças de configuração da ferramenta (modelo de prompt, testes e instruções) foram fundamentais para o aprimoramento da solução automatizada.

Considerando que o objetivo principal de uma auditoria preventiva em aquisições é “mitigar riscos que podem impactar os objetivos das futuras contratações” (BRASIL, 2000; CONTROLADORIA-GERAL DA UNIÃO, 2022), o uso de ferramenta baseada em LLM pode contribuir de forma direta para a eficiência dessas auditorias, identificando com alta confiabilidade os elementos dos artefatos de planejamento da contratação que estão em desconformidade com um determinado normativo, guiando, assim, a atenção dos auditores para pontos de maior risco de não conformidade.

Quanto às classes “Sim” e “Parcialmente”, o desempenho fraco ou mediano se deu, na maioria dos casos, porque a solução automatizada considerou parcial o atendimento dos ETPs aos comandos da Instrução Normativa SGD/ME nº 94/2022 quando os auditores consideraram essas situações como conformidade total. Tal situação representa um erro de baixa gravidade, uma vez que, no máximo, leva os auditores a analisarem com mais atenção pontos que eles, ao final da avaliação, podem considerar totalmente atendidos.

Além disso, a indicação de atendimento parcial em vez de total pelo modelo de linguagem parece ter ocorrido, basicamente, pelos seguintes motivos: (1) dificuldade de fazer com que o LLM faça uma análise contextualizada das necessidades e da solução que está sendo proposta nos ETPs, adequando-os à realidade do mercado e dos órgãos contratantes, e seja capaz de verificar a não aplicabilidade de um ou mais controles normativos dependendo do caso em questão; (2) dificuldade em garantir que a solução automatizada considere um determinado requisito normativo integralmente atendido a partir da análise conjugada de várias partes/seções dos ETPs; e (3) grande extensão dos requisitos normativos, que levam a ferramenta baseada em inteligência artificial a cobrar literalmente o atendimento ou menção a todos eles, sob pena de, na ausência de qualquer um, considerar a conformidade apenas como parcial.

O motivo (3) listado anteriormente chama a atenção pois, em casos como o do controle 4, nenhum dos 30 ETPs analisados fez referência a todos os itens exigidos pelo inciso II do art 11 da IN. Assim, os gestores responsáveis pela normatização do processo de contratações de TI devem analisar se o normativo, de fato, é exagerado em suas exigências, se os órgãos contratantes não estão atendendo a esses requisitos por falta de capacidade, ou mesmo se falta apenas uma melhor orientação aos órgãos contratantes para que se refiram a todos os itens exigidos, ainda que seja para indicar que eles não se aplicam ao caso concreto.

Finalmente, considera-se que a duração de 25,92 minutos apresentado pelo cenário 4 para conclusão da análise de 30 ETPs é extremamente satisfatória pois, em que pese não haver métricas para o desempenho humano, considerando a quantidade de informações que um Estudo Técnico Preliminar traz, bem como a extensão desse tipo de artefato, que pode chegar a dezenas de páginas, além da necessidade de redação da análise/conclusão fundamentada sobre a conformidade normativa para os vários controles, é razoável inferir que, nesse mesmo prazo de quase 26 minutos, um auditor concluiria a análise completa de apenas 1 ETP, o que representa um ganho de produtividade da ordem de 30 vezes.

5. Conclusão

O presente trabalho buscou avaliar se a utilização de grandes modelos de linguagem pode contribuir para o aumento da eficiência dos trabalhos de auditorias preventivas sobre processos de contratação de soluções de tecnologia da informação.

Os resultados apresentados demonstram que o uso de LLM contribui diretamente e com boa confiabilidade na indicação de pontos de desconformidade dos Estudos Técnicos Preliminares frente aos requisitos normativos da Instrução Normativa SGD/ME nº 94/2022. Além disso, propicia um ganho de produtividade elevado para as equipes de auditoria, viabilizando a análise completa de um ETP frente a esses requisitos em menos de 1 minuto.

Dessa forma, ainda que não substitua a análise humana, a ferramenta implementada neste estudo pode ser utilizada para auxiliar na identificação de potenciais irregularidades e na otimização do processo de auditoria, direcionando os auditores aos pontos de maior probabilidade de não conformidade e viabilizando tanto a aceleração quanto a ampliação da quantidade dos trabalhos de avaliação preventiva sobre contratações públicas de TI.

Destacou-se, para essa tarefa de análise automatizada, a importância dos modelos de prompt bem como das informações adicionais repassadas ao LLM para que ele tivesse o desempenho aprimorado.

Particularmente, o uso de técnicas de RAG mostrou-se eficiente para fornecer ao modelo de linguagem as informações mais específicas e relevantes sobre documentos extensos e sobre a análise que ele deveria realizar, resultando em avaliações bem estruturadas e fundamentadas.

Assim como toda ferramenta baseada em inteligência artificial, no entanto, é necessária uma revisão humana dos resultados apresentados pela solução automatizada com grandes modelos de linguagem, uma vez que ela apresenta alguns erros em suas avaliações.

Neste estudo, o erro mais comumente observado foi o julgamento de ETPs como estando em conformidade parcial quando, pela análise dos auditores, eles estariam em conformidade total. Trata-se de um erro de impacto relativamente baixo, no entanto, e que pode ser atribuído à dificuldade de interpretação contextualizada do normativo e dos artefatos de planejamento da contratação, bem como à extensão e à complexidade de alguns dos requisitos normativos.

Como limitações deste trabalho, destacam-se: (1) a inexistência de dados públicos de avaliações de conformidade normativa realizada por órgãos de auditoria; (2) a consequente necessidade de utilização de avaliações humanas realizadas por auditores de forma individual, e não institucional; (3) a quantidade relativamente baixa de ETPs analisados por auditores para fins de comparação com a solução automatizada; (4) a utilização de um único LLM; e (5) o fato de que, nas instruções para o modelo de linguagem, foram passados apenas alguns dos trechos mais relevantes dos documentos analisados, e não o conteúdo completo deles.

Possíveis trabalhos futuros podem explorar os mesmos cenários de automatização usando um ou mais modelos de linguagem diferentes do GPT-4o, utilizado neste estudo, a fim de verificar se as métricas de desempenho da solução se alteram. Podem ainda realizar as mesmas análises feitas aqui (com os mesmos prompts, testes e instruções), mas fornecendo o conteúdo completo dos documentos analisados. Por fim, pode-se aproveitar a mesma arquitetura computacional e expandir a análise para outros controles do mesmo normativo ou, ainda, aplicá-la a outro domínio de contratações públicas que não a área de tecnologia da informação.

Referências bibliográficas

- ARTIFICIAL ANALYSIS. **LLM API Provider Leaderboard**. [S. l.], 2024. Available at: <https://artificialanalysis.ai/leaderboards/providers>. Acesso em: 28 nov. 2024.
- BRASIL. **DECRETO No 3.591, DE 6 DE SETEMBRO 2000**. Dispõe sobre o Sistema de Controle Interno do Poder Executivo Federal e dá outras providências. Brasília: Presidência da República, 6 set. 2000.
- BRASIL. **INSTRUÇÃO NORMATIVA SGD/ME Nº 94, DE 23 DE DEZEMBRO DE 2022** Brasília: Ministério da Economia, 23 dez. 2022.
- BRASIL. **LEI Nº 14.133, DE 1º DE ABRIL DE 2021**. Lei de Licitações e Contratos Administrativos. Brasília: 1 abr. 2021.
- CASELO, H. de medeiros; NUNES, M. das G. V. **Processamento de Linguagem Natural - Conceitos, Técnicas e Aplicações em Português**. 2. ed. São Carlos: [s. n.], 2024.
- CHIANG, W.-L. Z. L. S. Y. N. A. A. L. T. L. D. Z. H. Z. B. J. M. E. G. J. S. I. **Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference**. [S. l.], 2024. Available at: <https://lmarena.ai/?leaderboard>. Acesso em: 28 nov. 2024.
- CHOWDHARY, K. R. **Fundamentals of Artificial Intelligence**. New Delhi: Springer, 2020. v. 37 Available at: https://doi.org/https://doi.org/10.1007/978-81-322-3972-7_19
- CHURCH, K. W. Word2Vec. **Natural Language Engineering**, [s. l.], v. 23, n. 1, p. 155–162, 2017. Available at: <https://doi.org/10.1017/S1351324916000334>
- CONTROLADORIA-GERAL DA UNIÃO. **Alice**. [S. l.], 2024a. Available at: <https://www.gov.br/cgu/pt-br/assuntos/auditoria-e-fiscalizacao/alice>. Acesso em: 26 nov. 2024.
- CONTROLADORIA-GERAL DA UNIÃO. **Orientação Prática: Serviços de Auditoria**. Brasília: [s. n.], 2022.
- CONTROLADORIA-GERAL DA UNIÃO. **Pesquisa - Relatórios de Auditoria da CGU**. [S. l.], 2024b. Available at: <https://eaud.cgu.gov.br/relatorios>. Acesso em: 27 set. 2024.
- DONG, C. *et al.* A Survey of Natural Language Generation. **ACM Computing Surveys**, [s. l.], v. 55, n. 8, 2022. Available at: <https://doi.org/10.1145/3554727>
- FACEBOOK. **FastText**. [S. l.], 2022. Available at: <https://fasttext.cc/>. Acesso em: 21 nov. 2024.
- FENG, S. Y. *et al.* A Survey of Data Augmentation Approaches for NLP. **Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021**, [s. l.], p. 968–988, 2021. Available at: <https://doi.org/10.18653/V1/2021.FINDINGS-ACL.84>
- HUGGING FACE. **Confusion Matrix**. [S. l.], 2024. Available at: https://huggingface.co/spaces/evaluate-metric/confusion_matrix. Acesso em: 20 nov. 2024.
- IZACARD, G. *et al.* Atlas: Few-shot Learning with Retrieval Augmented Language Models. [s. l.], 2022.
- LEWIS, P. *et al.* **Retrieval-Augmented Generation for Knowledge-Intensive NLP**

Tasks. [S. l.: s. n.], 2021.

MINAEE, S. *et al.* Large Language Models: A Survey. [s. l.], 2024.

MUHLGAY, D. *et al.* Generating Benchmarks for Factuality Evaluation of Language Models. **EACL 2024 - 18th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference**, [s. l.], v. 1, p. 49–66, 2023.

OECD. **GENERATIVE AI FOR ANTI-CORRUPTION AND INTEGRITY IN GOVERNMENT - TAKING STOCK OF PROMISE, PERILS AND PRACTICE.** [S. l.: s. n.], 2024.

OPENAI *et al.* GPT-4 Technical Report. [s. l.], 2023.

PAIVA, E. S. de *et al.* Continued pre-training of LLMs for Portuguese and Government domain: A proposal for product identification in textual purchase descriptions. [s. l.], 2024.

PATWARDHAN, N.; MARRONE, S.; SANSONE, C. **Transformers in the Real World: A Survey on NLP Applications.** [S. l.]: MDPI, 2023. Available at: <https://doi.org/10.3390/info14040242>

PENNINGTON, J.; SOCHER, R.; MANNING, C. D. GloVe: Global Vectors for Word Representation. *In:* , 2014. **Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)**. [S. l.: s. n.], 2014. p. 1532–1543.

RAM, O. *et al.* In-Context Retrieval-Augmented Language Models. **Transactions of the Association for Computational Linguistics**, [s. l.], v. 11, p. 1316–1331, 2023. Available at: https://doi.org/https://doi.org/10.1162/tacl_a_00605

SARKER, I. H. LLM potentiality and awareness: a position paper from the perspective of trustworthy and responsible AI modeling. **Discover Artificial Intelligence**, [s. l.], v. 4, n. 1, 2024. Available at: <https://doi.org/10.1007/s44163-024-00129-0>

SCHULHOFF, S. *et al.* The Prompt Report: A Systematic Survey of Prompting Techniques. [s. l.], 2024.

SCIKIT LEARN. **Confusion_Matrix.** [S. l.], 2024. Available at: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.confusion_matrix.html. Acesso em: 20 nov. 2024.

SOBRINO-GARCÍA, I. Artificial Intelligence Risks and Challenges in the Spanish Public Administration: An Exploratory Analysis through Expert Judgements. **Administrative Sciences 2021, Vol. 11, Page 102**, [s. l.], v. 11, n. 3, p. 102, 2021. Available at: <https://doi.org/10.3390/ADMSCI11030102>

TOLEDO, A. T. de; MENDONÇA, M. A aplicação da inteligência artificial na busca de eficiência pela administração pública. **Revista do Serviço Público - RSP**, [s. l.], v. 74, n. 2, p. 410–438, 2023. Available at: <https://doi.org/10.21874/rsp.v74i2.6829>

WANG, Z. *et al.* Interactive Natural Language Processing. [s. l.], 2023.

YOUNG, M. M. *et al.* Artificial Intelligence and Administrative Evil. **Perspectives on Public Management and Governance**, [s. l.], v. 4, n. 3, p. 244–258, 2021. Available at: <https://doi.org/10.1093/PPMGOV/GVAB006>

ZHANG, Y. *et al.* Pushing The Limit of LLM Capacity for Text Classification. [s. l.], 2024.

ZHAO, W. X. *et al.* A Survey of Large Language Models. [s. l.], 2023.

Apêndice 1 – Controles da IN SGD/ME nº 94/2022 sobre Estudos Técnicos Preliminares

Com base nos requisitos previstos pela Instrução Normativa SGD/ME nº 94/2022 em sua subseção II – Do Estudo Técnico Preliminar, elaborou-se o checklist apresentado na Tabela 8, com os testes de conformidade que deveriam ser aplicados sobre os ETPs de modo a verificar a conformidade desses documentos em relação ao normativo disciplinador das contratações de TIC no Poder Executivo Federal.

A análise de conformidade de um determinado ETP deve produzir uma resposta dividida em 2 partes: a inicial, na qual se indica, na coluna “Conformidade” do checklist, se o documento atende plenamente aos requisitos normativos, podendo a resposta ser “Sim”, “Não”, “Parcialmente” ou “Não se Aplica”; e a parte final, em que deve ser apresentado o embasamento da parte inicial da resposta com base nos trechos do documento analisado e no que é solicitado pelo normativo.

Tabela 8 - Checklist de testes de auditoria sobre ETPs baseado na IN SGD/ME nº 94/2022

Estudos Técnicos Preliminares	Teste	Descrição	Critério	Conformidade	Explicação
	1	Verificar no ETP se foram definidas as necessidades de negócio e tecnológicas e os requisitos necessários e suficientes à escolha da solução de TIC.	Art. 11, Inc. I, da IN SGD/ME nº 94/2022; e Art. 18, §1º, Inciso III, da Lei nº 14.133/21		
	2	Verificar no ETP se a solução tecnológica escolhida (objeto da contratação) resolve o problema do órgão ou entidade e/ou atende à necessidade descrita no DFD.	Art. 10, da IN SGD/ME nº 94/2022; e Art. 18, §1º, Incisos I e XIII, da Lei nº 14.133/21		
	3	Verificar se consta no ETP, de forma <i>detalhada, motivada e justificada</i> , inclusive quanto à forma de cálculo, o quantitativo de bens e serviços necessários para a composição da solução de TIC.	Art. 11, Inc. I, da IN SGD/ME nº 94/2022; e Art. 18, §1º, Inciso IV, da Lei nº 14.133/21		
	4	Verificar no ETP se foi realizada análise comparativa de soluções , observando os aspectos <i>econômicos e qualitativos</i> em termos de benefício para o alcance dos objetivos da contratação. Obs.: quando couber, observar se tal comparativo considerou as diferentes métricas de pagamento ou formas de remuneração possíveis (Ex.: aluguel versus compra; preço global versus unitário etc.).	Art. 11, Inc. II, da IN SGD/ME nº 94/2022; e Art. 18, §1º, Inciso V, da Lei nº 14.133/21		
	5	Verificar no ETP se foi realizada a análise comparativa de custos, por meio da comparação de custos totais de propriedade (Total Cost Ownership - TCO) das soluções consideradas viáveis.	Art. 11, Inc. III, da IN SGD/ME nº 94/2022		
	6	Verificar no ETP se a forma de pagamento (remuneração) escolhida encontra-se adequadamente fundamentada conforme critérios técnicos e financeiros.	Art. 18, Inc. IV, da IN SGD/ME nº 94/2022; e Súmula TCU nº 269		
	7	Caso se trate de procedimento de adesão a ata de registro de preços, verificar se no ETP consta registro do ganho de eficiência, da viabilidade e da economicidade para a administração pública federal da utilização da ata de registro de preços.	Art. 9º, §3º, da IN SGD/ME nº 94/2022; e art. 86, §2º, da Lei nº 14.133/21		

Fonte: Elaborado pelo autor

Apêndice 2 – Modelos de Prompt, Testes (perguntas) e Instruções

A seguir, são apresentados os modelos de prompt, as perguntas utilizadas para os testes de auditoria, e as instruções repassadas ao LLM para que ele realizasse a análise de conformidade dos Estudos Técnicos Preliminares em relação aos controles previstos na Instrução Normativa SGD/ME nº 94/2022.

Tabela 9 - Modelos de prompt utilizados

Modelo de prompt PPT_01	Modelo de prompt PPT_02
<p>""Você é um auditor interno do governo federal brasileiro, especializado em contratações públicas de soluções de tecnologia da informação. Você fará uma análise de conformidade legal, verificando se um determinado artefato de planejamento de contratação atende ao que é exigido em um normativo.</p> <p>O quesito/controle do normativo a ser atendido será passado como pergunta, e, como contexto, serão passados os trechos mais relevantes do normativo e do artefato a ser analisado.</p> <p>## Pergunta: {{query}}</p> <p>## Contexto do normativo: {{ctx_in94}}</p> <p>## Trechos do artefato: {{trechos_artefato}}</p> <p>##Resposta: Você deve iniciar com sua resposta com "Sim.", "Não." ou "Parcialmente." e, em seguida, apresentar o embasamento de sua resposta com base no contexto dos documentos fornecidos, incluindo o número dos artigos do normativo que sustentam sua análise.</p> <p>Caso o texto do artefato apenas repita o texto do normativo, deve-se considerar que o documento não atende ao requisito, fornecendo a resposta no padrão solicitado anteriormente.</p> <p>""</p>	<p>""Você é um auditor interno do governo federal brasileiro, especializado em contratações públicas de soluções de tecnologia da informação. Você fará uma análise de conformidade legal, verificando se um determinado artefato de planejamento de contratação atende ao que é exigido em um normativo.</p> <p>O quesito/controle do normativo a ser atendido será passado como pergunta, e, como contexto, serão passados os trechos mais relevantes do normativo, instruções de como eles devem ser avaliados, além dos trechos mais relevantes do artefato a ser analisado.</p> <p>## Pergunta: {{query}}</p> <p>## Contexto do normativo: {{ctx_in94}}</p> <p>## Instruções de análise do controle normativo: {{instrucao_controle}}</p> <p>## Trechos do artefato: {{trechos_artefato}}</p> <p>##Resposta: Você deve iniciar com sua resposta com "Sim.", "Não." ou "Parcialmente." e, em seguida, apresentar o embasamento de sua resposta com base no contexto dos documentos fornecidos, incluindo o número dos artigos do normativo que sustentam sua análise.</p>

Modelo de prompt PPT_01	Modelo de prompt PPT_02
	Caso o texto do artefato apenas repita o texto do normativo, deve-se considerar que o documento não atende ao requisito, fornecendo a resposta no padrão solicitado anteriormente.
	""""

Fonte: Elaborado pelo autor

Tabela 10 - Perguntas utilizadas para testes de auditoria realizados

Teste	Conjunto de perguntas PGT_01	Conjunto de perguntas PGT_02
1	"As necessidades de negócio e tecnológicas e os requisitos necessários e suficientes à escolha da solução de TIC foram definidos?"	"As necessidades de negócio e tecnológicas e os requisitos necessários e suficientes à escolha da solução de TIC foram definidos? Observe o que está previsto no Artigo 11, inciso I, da IN SGD/ME nº 94/2022.",
2	"A solução tecnológica escolhida (objeto da contratação) resolve o problema do órgão ou entidade e/ou atende à necessidade descrita no Documento de Formalização da Demanda (DFD)?"	"A solução tecnológica escolhida (objeto da contratação) resolve o problema do órgão ou entidade e/ou atende à necessidade descrita no Documento de Formalização da Demanda (DFD)?"
3	"Consta do Estudo Técnico Preliminar, de forma detalhada, motivada e justificada, inclusive quanto à forma de cálculo, o quantitativo de bens e serviços necessários para a composição da solução de TIC?"	"Consta do Estudo Técnico Preliminar, de forma detalhada, motivada e justificada, inclusive quanto à forma de cálculo, o quantitativo de bens e serviços necessários para a composição da solução de TIC? Observe o que está previsto no artigo 11, inciso I, e artigo 14 da IN SGD/ME 94/2022"
4	"No Estudo Técnico Preliminar, foi realizada análise comparativa de soluções, observando os aspectos econômicos e qualitativos em termos de benefício para o alcance dos objetivos da contratação?"	"No Estudo Técnico Preliminar, foi realizada análise comparativa de soluções, observando os aspectos econômicos e qualitativos em termos de benefício para o alcance dos objetivos da contratação? Observe o disposto no Artigo 11, inciso II da IN SGD/ME nº 94/2022."
5	"No Estudo Técnico Preliminar, foi realizada a análise comparativa de custos, por meio da comparação de custos totais de propriedade (Total Cost Ownership - TCO) das soluções consideradas viáveis?"	"No Estudo Técnico Preliminar, foi realizada a análise comparativa de custos, por meio da comparação de custos totais de propriedade (Total Cost Ownership - TCO) das soluções consideradas viáveis? Observe o disposto no Artigo 11, inciso III da IN SGD/ME nº 94/2022."
6	"No Estudo Técnico Preliminar, a forma de pagamento (remuneração) escolhida encontra-se adequadamente fundamentada conforme critérios técnicos e financeiros?"	"No Estudo Técnico Preliminar, a forma de pagamento (remuneração) escolhida encontra-se adequadamente fundamentada conforme critérios técnicos e financeiros? Observe o que está previsto no Artigo 18, Inc. IV, da IN SGD/ME nº 94/2022."
7	"Caso o Estudo Técnico Preliminar (ETP) esteja propondo a contratação por meio da adesão a Sistema de Registro de Preços (SRP) ou Ata de Registro de Preços (ARP), consta desse ETP registro do ganho de eficiência, da viabilidade e da economicidade para a administração pública federal da utilização da ata ou do sistema de registro de preços? Caso não se esteja prevendo a contratação por meio de ARP ou SRP, responda 'N/a'."	"Caso o Estudo Técnico Preliminar (ETP) esteja propondo a contratação por meio da adesão a Sistema de Registro de Preços (SRP) ou Ata de Registro de Preços (ARP), consta desse ETP registro do ganho de eficiência, da viabilidade e da economicidade para a administração pública federal da utilização da ata ou do sistema de registro de preços? Caso não se esteja prevendo a contratação por meio de ARP ou SRP, responda 'N/A'. Observe o disposto no Artigo 9, parágrafo 3º da IN SGD/ME nº 94/2022."

Fonte: Elaborado pelo autor

Tabela 11 - Conjunto de instruções INST_01 para os testes de auditoria realizados

Teste	Conjunto de Instruções INST_01
1	"Se forem apresentadas apenas necessidades genéricas, que qualquer organização tem em relação a tecnologia da informação, o atendimento ao requisito é apenas parcial. No entanto, se forem apresentadas necessidade tecnológicas e de negócio específicas, mesmo que poucas e ainda que espalhadas ao longo de várias seções do artefato, deve-se considerar o controle como integralmente atendido (resposta afirmativa para a pergunta)."
2	"Se as necessidades não tiverem sido descritas adequadamente (resposta 'Não' ou 'Parcialmente' na pergunta anterior), a resposta aqui não pode ser positiva (podendo ser 'Não' ou 'Parcialmente'). Além disso, deve estar claro qual é a solução (o objeto) a ser contratado. Então, caso seja demonstrado, ao longo do artefato, que a solução escolhida atende às necessidades elencadas no próprio artefato, ainda que não haja menção ao DFD, deve-se considerar o requisito como atendido (resposta 'Sim' à pergunta). Por fim, a análise deve levar em conta apenas a solução escolhida, ainda que ela não tenha sido comparada a outras."
3	"Observe que a mera apresentação de quantidades finais de itens/serviços a serem contratados não atende a esse requisito. Se não houver uma explicação matemática de como foram definidos, a partir das necessidades, os quantitativos finais de serviço ou itens a serem contratados, a resposta deve ser negativa. Se houver explicação sobre como foi determinada a quantidade a ser contratada apenas para parte dos itens que compõem a solução escolhida, o atendimento a esse critério será parcial (resposta 'Parcialmente'). Além disso, a quantidade de bens/serviços a ser contratada deve constar do ETP, não sendo adequado indicar que ela se encontra em outros documentos, como Documento de Formalização da Demanda (DFD) ou Termo de Referência (TR), já que ela é informação crucial para a tomada da decisão final do ETP, o qual antecede a etapa de elaboração do TR."
4	"Quando couber, observar se tal comparativo considerou as diferentes métricas de pagamento ou formas de remuneração possíveis (Ex.: aluguel versus compra; preço global versus unitário etc.)."
5	"A mera menção de foi feita essa análise, ou a existência de uma seção no artefato para isso, não significa que a análise de TCO foi realizada. Você deve verificar o conteúdo dos trechos do artefato. Deve haver uma análise de custos indiretos associados às soluções viáveis. Esses custos indiretos são aqueles inerentes ao ciclo de vida dos bens e serviços de cada solução, incluindo valores de aquisição, insumos, garantia técnica estendida, manutenção, migração e treinamento. Se há uma análise de pelo menos parte desses custos indiretos, considera-se que a análise foi feita. Se não há qualquer análise de custos indiretos, considera-se que a análise de TCO não foi realizada. Por fim, essa análise deve ser feita apenas para as soluções indicadas como viáveis. Se há apenas uma solução identificada como viável, não há problema e esse quesito pode ser considerado atendido caso seja feita a análise de custos indiretos para a solução."
6	"A mera definição do custo da solução ou realização de pesquisa de preços ou de mercado não atende a esse requisito nem parcialmente (resposta negativa para a pergunta). É necessário indicar e explicar de que forma ocorrerá o pagamento pela solução. Por exemplo: pagamento único ou parcelado em etapas; global ou por item; por entrega ou por posto de trabalho. Em suma: deve-se explicar como o objeto deve ser entregue para que haja o pagamento e que parcela do pagamento será feita a cada entrega."
7	"Alegações genéricas de aumento de eficiência ou da economicidade, de redução de preços ou economia de escala, ou ainda de economia de tempo, sem o detalhamento dessas alegações para o caso específico do órgão e da solução que ele pretende adquirir não atendem ao requisito."

Fonte: elaborado pelo autor

Tabela 12 - Conjunto de instruções INST_02 para os testes de auditoria

Teste	INST_02
1	<p>""1) Se forem apresentadas APENAS necessidades genéricas, que qualquer organização tem em relação a tecnologia da informação, o requisito deve ser considerado não atendido. No entanto, se forem apresentadas necessidade tecnológicas e de negócio específicas, mesmo que poucas e ainda que espalhadas ao longo de várias seções do artefato, deve-se considerar o controle como integralmente atendido.</p> <p>2) Não seja tão rigoroso! Não é necessário especificar sempre necessidades de capacitação, legais, temporais, sociais, ambientais e culturais, conforme descrito no artigo 16, inciso I, da IN SGD/ME nº 94/2022. Assim, se elas não forem descritas, mas houver descrição de necessidades de negócio e tecnológicas, pode-se considerar o quesito como atendido.</p> <p>3) Se forem apresentadas necessidades de negócio ou tecnológicas específicas do contratante, não julgue se eles são limitados. Considere que eles são suficientes e prossiga com a análise.</p> <p>4) Para este quesito/pergunta, não é necessário definir quantitativos de bens e serviços a serem contratados. Deve-se focar apenas na definição de necessidades/requisitos, conforme explicado anteriormente.</p> <p>5) Por fim, a indicação do que se pretende contratar não atende ao requisito, pois já representa a solução escolhida, e não as necessidades a serem atendidas.""</p>
2	<p>""1) Se as necessidades de negócio e tecnológicas não tiverem sido descritas adequadamente ou forem genéricas (resposta 'Não' ou 'Parcialmente' na pergunta anterior), a resposta aqui não pode ser positiva (podendo ser apenas 'Não' ou 'Parcialmente').</p> <p>2) Além disso, deve estar clara qual é a solução (o objeto) a ser contratado. Então, caso seja demonstrado, ao longo do artefato, que a solução escolhida atende às necessidades elencadas no próprio artefato, ainda que não haja menção ao DFD, deve-se considerar o requisito como atendido.</p> <p>3) Além disso, a análise deve levar em conta apenas a solução escolhida, ainda que ela não tenha sido comparada a outras.</p> <p>4) A indicação de benefícios alcançados com a contratação da solução escolhida, por si só, não atende ao quesito normativo. É preciso demonstrar que a solução atende às necessidades elencadas no artefato ou no DFD.""</p>
3	<p>""1) Observe que a mera apresentação de quantidades finais de itens/serviços a serem contratados não atende a esse requisito.</p> <p>2) Se não houver uma explicação de como foram definidos, a partir das necessidades, os quantitativos finais de serviço ou itens a serem contratados, a resposta deve ser negativa.</p> <p>3) Se houver explicação sobre como foi determinada a quantidade a ser contratada apenas para parte dos itens que compõem a solução escolhida, o atendimento a esse critério será parcial (resposta 'Parcialmente').</p> <p>4) A explicação sobre como foram determinados os quantitativos pode ser baseado no histórico de consumo do mesmo serviço ou produto.</p> <p>5) Além disso, a quantidade de bens/serviços a ser contratada deve constar obrigatoriamente do ETP, não sendo adequado indicar que ela se encontra no Termo de Referência (TR), já que ela é informação crucial para a tomada da decisão final do ETP, o qual antecede a etapa de elaboração do TR.""</p>
4	<p>""1) Quando couber, observar se tal comparativo considerou as diferentes métricas de pagamento ou formas de remuneração possíveis (Ex.: aluguel versus compra; preço global versus unitário etc.)</p> <p>2) Não seja tão rigoroso! Se o documento afirma haver apenas uma solução que atende às necessidades do órgão contratante, assuma isso como verdadeiro. Assim, se a análise para essa solução for feita de forma adequada, considera-se o quesito normativo atendido.</p> <p>3) No entanto, cenários em que uma das soluções é a aquisição de equipamentos sempre devem considerar, pelo menos, a opção de aluguel ou outsourcing em vez da aquisição.</p> <p>4) No caso de necessidades de substituição ou aquisição de equipamentos, a pesquisa de diferentes configurações / fabricantes de equipamentos não representa uma análise comparativa de soluções, uma vez que, nesse caso, já está sendo definida como solução a "aquisição" de equipamentos de determinado tipo (ex: desktops). Uma análise comparativa deve levantar alternativas justamente à aquisição, como aluguel, outsourcing ou mesmo a aquisição de outros tipos de equipamentos. Exemplo: notebooks em vez de desktops.</p> <p>5) A mera estimativa de valor da contratação ou a metodologia utilizada para o cálculo do preço de referência não constituem uma análise comparativa de soluções</p>

Teste

INST_02

6) Detalhes sobre a forma de contratação pública (ex: contratação direta, realização de licitação / pregão eletrônico ou uso de Sistema de Registro de Preço) não fazem parte da "solução" e, portanto, não devem ser consideradas para fins de análise desse quesito normativo.

7) A simples menção a uma análise de contratações similares feitas por outros órgãos não atende ao requisito normativo. Se não for apresentada ao menos uma solução concreta contratada por outro órgão ou indicado que não foram encontradas alternativas similares em outros órgãos, considera-se que o requisito normativo não foi atendido.

8) A justificativa para o parcelamento ou não da solução NÃO faz parte da análise comparativa de soluções. Trata-se de consideração posterior, que deve ser feita já sobre a solução escolhida.

9) Se, como análise comparativa de soluções, o artefato analisado indicar apenas pesquisa de preços ou a justificativa para o parcelamento da contratação, considera-se que o quesito normativo não foi atendido."""

""1) Para que esse quesito normativo seja considerado como plenamente atendido, deve haver uma análise de custos indiretos associados às soluções viáveis. Esses custos indiretos são aqueles inerentes ao ciclo de vida dos bens e serviços de cada solução, incluindo valores de aquisição, insumos, garantia técnica estendida, manutenção, migração e treinamento.

Se há uma análise de pelo menos parte desses custos indiretos, considera-se que a análise foi feita. Se não há qualquer análise de custos indiretos, considera-se que a análise de TCO não foi realizada (resposta negativa para a pergunta).

5 2) A mera menção de foi feita essa análise, ou a existência de uma seção no artefato para isso, não significa que a análise de TCO foi realizada. Você deve verificar o conteúdo dos trechos do artefato.

3) A simples projeção do custo/valor/preço da solução ao longo dos anos de um potencial contrato não representa uma análise de TCO.

4) A análise de TCO deve ser feita apenas para as soluções indicadas como viáveis. Se há apenas uma solução identificada como viável, não há problema e esse quesito pode ser considerado atendido caso seja feita a análise de custos indiretos para a solução.

5) Custos obtidos por pesquisa de preço/mercado, que representam o quanto o órgão contratante pagará diretamente à contratada pela solução não contemplam os custos indiretos e, assim, não representam uma análise de TCO."""

""1) A mera definição do custo da solução ou realização de pesquisa de preços ou de mercado não atende a esse requisito nem parcialmente (resposta negativa para a pergunta). É necessário indicar e explicar de que forma ocorrerá o pagamento pela solução. Por exemplo: pagamento único ou parcelado em etapas; global ou por item; por entrega ou por posto de trabalho. Em suma: deve-se explicar como o objeto deve ser entregue para que haja o pagamento e que parcela do pagamento será feita a cada entrega.

6 2) A definição da métrica de medição dos serviços (ex: Unidade de Serviços de Nuvem (USN), Unidade de Serviço Técnico (UST), Ponto de Função (PF), entre outras) deve ser considerada parte da análise sobre forma de remuneração. Assim, se a escolha dessa métrica for explicada, pode-se considerar o quesito normativo como atendido.

3) Argumentos sobre a viabilidade ou não do parcelamento da solução não fazem parte das considerações sobre a forma de remuneração e, portanto, não atendem a esse quesito normativo."""

7 ""1) Alegações genéricas de aumento de eficiência ou da economicidade, de redução de preços ou economia de escala, ou ainda de economia de tempo, sem o detalhamento dessas alegações para o caso específico do órgão e da solução que ele pretende adquirir não atendem ao requisito."""

Tabela 13 - Conjunto de instruções INST_03 para os testes de auditoria realizados

Teste	Conjunto de Instruções INST_03
1	<p>""1) Se forem apresentadas APENAS necessidades genéricas, que qualquer organização tem em relação a tecnologia da informação, o requisito deve ser considerado não atendido (nem parcialmente).</p> <p>2) A indicação de necessidades e requisitos genéricos que qualquer organização, pública ou privada, apresenta, como "aumento de eficiência", "aumento de qualidade", "aumento de economicidade" também não atende a esse requisito, nem mesmo parcialmente.</p> <p>3) No entanto, se forem apresentadas necessidade tecnológicas E de negócio específicas dos órgãos contratantes, mesmo que poucas e ainda que espalhadas ao longo de várias seções do artefato, deve-se considerar o controle como integralmente atendido.</p> <p>4) Não seja tão rigoroso! Não é necessário especificar sempre necessidades de capacitação, legais, temporais, sociais, ambientais e culturais, conforme descrito no artigo 16, inciso I, da IN SGD/ME nº 94/2022. Assim, se elas não forem descritas, mas houver descrição de necessidades de negócio e tecnológicas, pode-se considerar o quesito como atendido.</p> <p>5) Se forem apresentadas necessidades de negócio ou tecnológicas específicas do contratante, não julgue se eles são limitados. Considere que eles são suficientes e prossiga com a análise.</p> <p>6) Para este quesito/pergunta, não é necessário definir quantitativos de bens e serviços a serem contratados. Deve-se focar apenas na definição de necessidades/requisitos, conforme explicado anteriormente.</p> <p>7) Por fim, a indicação dos itens ou serviços que se pretende contratar ou de suas quantidades não atende a esse quesito normativo, pois já representa a solução escolhida, e não as necessidades a serem atendidas.""</p>
2	<p>""1) Para considerar esse quesito normativo como atendido, é necessário que o artefato demonstre que a solução escolhida atende às necessidades elencadas no próprio artefato ou no Documento de Formalização da Demanda (DFD).</p> <p>2) Assim, se as necessidades de negócio e tecnológicas não tiverem sido descritas adequadamente ou forem genéricas a resposta aqui não pode ser positiva, podendo ser apenas 'Não' ou 'Parcialmente'.</p> <p>3) Se não há uma relação entre a solução escolhida e as necessidades específicas da instituição contratante, o quesito normativo não será atendido.</p> <p>4) Além disso, deve estar clara qual é a solução (o objeto) escolhida. Se isso não estiver claro (por exemplo, por existência de mais de uma solução possível sem escolha de uma delas), o quesito normativo também não será atendido.</p> <p>5) A análise deve levar em conta apenas a solução 'escolhida', ainda que ela não tenha sido comparada a outras.</p> <p>6) A indicação de benefícios alcançados com a contratação da solução escolhida, por si só, não atende ao quesito normativo. É preciso demonstrar que a solução atende às necessidades elencadas no artefato ou no DFD.</p> <p>7) A definição da forma de contratação da solução não se confunde com a solução. Assim, se o órgão menciona uma determinada forma de contratação, como Pregão Eletrônico ou mesmo o Sistema de Registro de Preços (SRP), e não o objeto a ser contratado, para indicar que ela atende às necessidades do órgão, o quesito normativo não está atendido.</p> <p>8) A descrição da solução e seus quantitativos, por si só, não atende a esse requisito, pois o que deve ser apresentado é a forma como esses itens resolvem os problemas/necessidades elencados no artefato.""</p>
3	<p>""1) Observe que a mera apresentação de quantidades finais ou da estimativa de preço de itens/serviços a serem contratados não atende a esse requisito.</p> <p>2) Se não houver uma explicação matemática (memória de cálculo) de como foram definidos, a partir das necessidades, os quantitativos finais de serviço ou itens a serem contratados, a resposta deve ser negativa.</p> <p>3) Se houver explicação sobre como foi determinada a quantidade a ser contratada apenas para parte dos itens que compõem a solução escolhida, o atendimento a esse critério será parcial (resposta 'Parcialmente').</p>

Teste

Conjunto de Instruções INST_03

4) Além disso, a quantidade de bens/serviços a ser contratada deve constar do ETP, não sendo adequado indicar que ela se encontra em outros documentos, como Documento de Formalização da Demanda (DFD) ou Termo de Referência (TR), já que ela é informação crucial para a tomada da decisão final do ETP, o qual antecede a etapa de elaboração do TR."""

""1) Quando couber, observar se tal comparativo considerou as diferentes métricas de pagamento ou formas de remuneração possíveis (Ex.: aluguel versus compra; preço global versus unitário etc.)

2) Não seja tão rigoroso! Se o documento afirma haver apenas uma solução que atende às necessidades do órgão contratante, assuma isso como verdadeiro. Assim, se a análise para essa solução for feita de forma adequada, considera-se o quesito normativo atendido.

3) No entanto, cenários em que uma das soluções é a aquisição de equipamentos sempre devem considerar, pelo menos, a opção de aluguel ou outsourcing em vez da aquisição.

4) No caso de necessidades de substituição ou aquisição de equipamentos, a pesquisa de diferentes configurações / fabricantes de equipamentos não representa uma análise comparativa de soluções, uma vez que, nesse caso, já está sendo definida como solução a "aquisição" de equipamentos de determinado tipo (ex: desktops). Uma análise comparativa deve levantar alternativas justamente à aquisição, como aluguel, outsourcing ou mesmo a aquisição de outros tipos de equipamentos. Exemplo: notebooks em vez de desktops.

5) A mera estimativa de valor da contratação ou a metodologia utilizada para o cálculo do preço de referência não constituem uma análise comparativa de soluções

6) Detalhes sobre a forma de contratação pública (ex: contratação direta, realização de licitação / pregão eletrônico ou uso de Sistema de Registro de Preço) não fazem parte da "solução" e, portanto, não devem ser consideradas para fins de análise desse quesito normativo.

7) A simples menção a uma análise de contratações similares feitas por outros órgãos não atende ao requisito normativo. Se não for apresentada ao menos uma solução concreta contratada por outro órgão ou indicado que não foram encontradas alternativas similares em outros órgãos, considera-se que o requisito normativo não foi atendido.

8) A justificativa para o parcelamento ou não da solução NÃO faz parte da análise comparativa de soluções. Trata-se de consideração posterior, que deve ser feita já sobre a solução escolhida.

9) Se, como análise comparativa de soluções, o artefato analisado indicar apenas pesquisa de preços ou a justificativa para o parcelamento da contratação, considera-se que o quesito normativo não foi atendido.""",

""1) Para que esse quesito normativo seja considerado como plenamente atendido, deve haver uma análise de custos indiretos associados às soluções viáveis. Esses custos indiretos são aqueles inerentes ao ciclo de vida dos bens e serviços de cada solução, incluindo valores de aquisição, insumos, garantia técnica estendida, manutenção, migração e treinamento. Se há uma análise de pelo menos parte desses custos indiretos, considera-se que a análise foi feita. Se não há qualquer análise de custos indiretos, considera-se que a análise de TCO não foi realizada (resposta negativa para a pergunta).

2) A mera menção de foi feita essa análise, ou a existência de uma seção no artefato para isso, não significa que a análise de TCO foi realizada. Você deve verificar o conteúdo dos trechos do artefato.

3) A simples projeção do custo/valor/preço da solução ao longo dos anos de um potencial contrato não representa uma análise de TCO.

4) A análise de TCO deve ser feita apenas para as soluções indicadas como viáveis. Se há apenas uma solução identificada como viável, não há problema e esse quesito pode ser considerado atendido caso seja feita a análise de custos indiretos para a solução.

5) Custos obtidos por pesquisa de preço/mercado, que representam o quanto o órgão contratante pagará diretamente à contratada pela solução não contemplam os custos indiretos e, assim, não representam uma análise de TCO.""

""1) A mera definição do custo da solução ou realização de pesquisa de preços ou de mercado não atende a esse requisito nem parcialmente (resposta negativa para a pergunta). É necessário indicar e explicar de que forma ocorrerá o pagamento pela solução.

6) Por exemplo: pagamento único ou parcelado em etapas; global ou por item; por entrega ou por posto de trabalho. Em suma: deve-se explicar como o objeto deve ser entregue para que haja o pagamento e que parcela do pagamento será feita a cada entrega.

Teste

Conjunto de Instruções INST_03

2) A definição da métrica de medição dos serviços (ex: Unidade de Serviços de Nuvem (USN), Unidade de Serviço Técnico (UST), Ponto de Função (PF), entre outras) deve ser considerada parte da análise sobre forma de remuneração.

Assim, se a escolha dessa métrica for explicada, pode-se considerar o quesito normativo como atendido.

3) Argumentos sobre a viabilidade ou não do parcelamento da solução não fazem parte das considerações sobre a forma de remuneração e, portanto, não atendem a esse quesito normativo."""

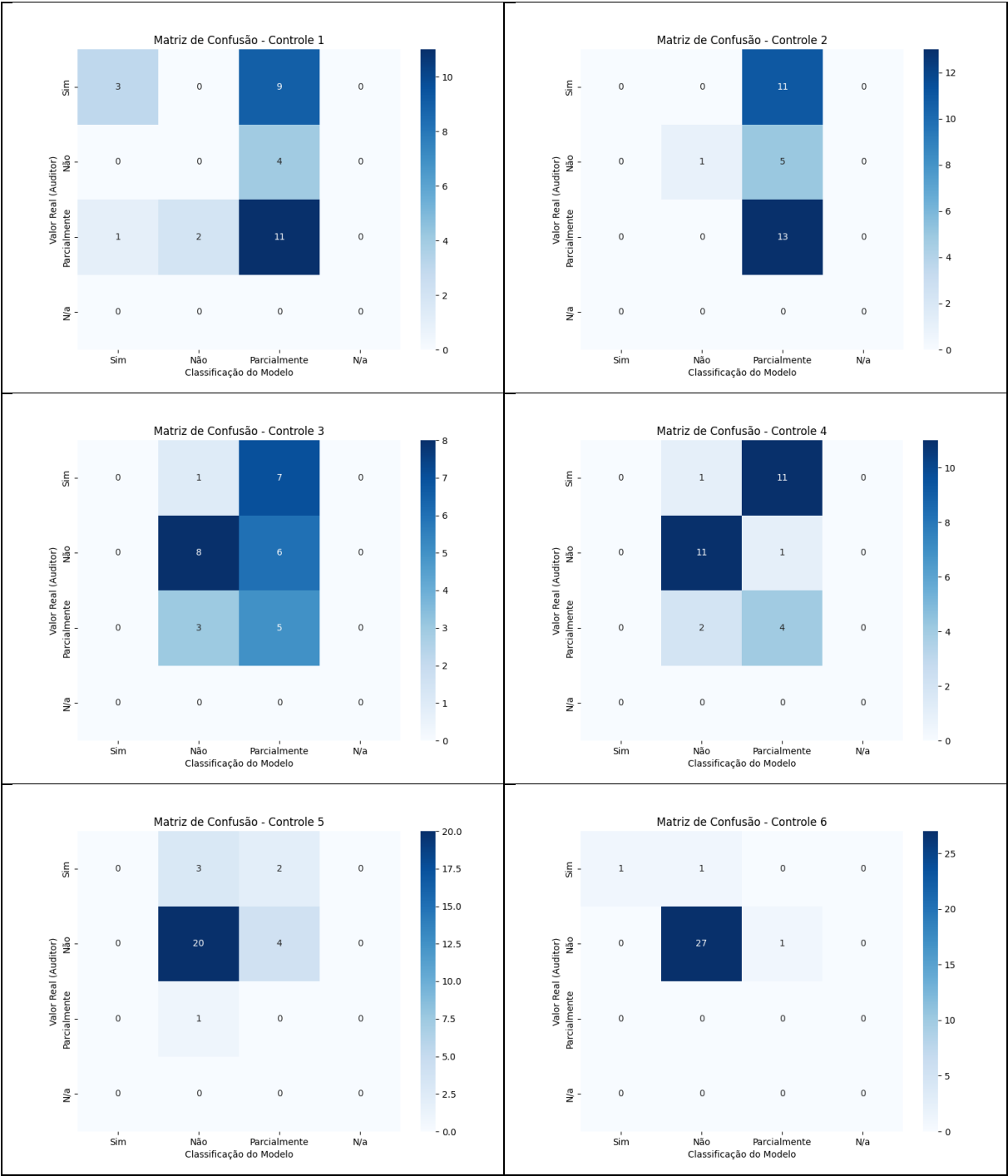
7 ""1) Alegações genéricas de aumento de eficiência ou da economicidade, de redução de preços ou economia de escala, ou ainda de economia de tempo, sem o detalhamento dessas alegações para o caso específico do órgão e da solução que ele pretende adquirir não atendem ao requisito."""

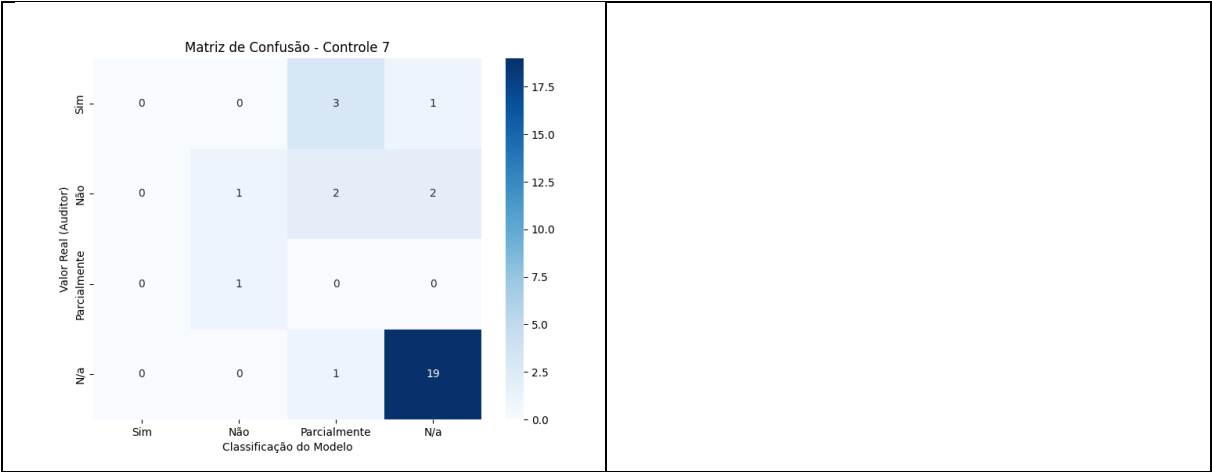
Fonte: Elaborado pelo autor

Apêndice 3 – Matrizes de Confusão para o Cenário 4

As matrizes de confusão geradas para as análises do Cenário 4 são apresentadas no Quadro 1.

Quadro 1 - Matrizes de confusão do cenário 4 para os 7 testes de auditoria





Fonte: elaborado pelo autor

No eixo vertical, estão as classes de referência, indicadas pelos auditores. No eixo horizontal, aquelas que foram previstas pelo LLM. Assim, as células da diagonal principal da matriz indicam a quantidade de acertos da classificação pelo modelo de linguagem (classe prevista = classe real). Já as quantidades das demais células indicam o número de erros de um determinado tipo. Por exemplo, no teste 1, entre as 4 previsões de “Sim” feitas pelo modelo, 3 estavam corretas (precisão de 0,75) e 1 errada (de acordo com os auditores, o correto seria “Parcialmente”).



Missão

Aprimorar a Administração Pública
em benefício da sociedade por meio
do controle externo

Visão

Ser referência na promoção de uma
Administração Pública efetiva, ética,
ágil e responsável