
**Aprimorando o atendimento ao cidadão com
RAGLC (Retrieval Augmented Generation in
Logic Context), uma nova abordagem RAG**

Abel Melo Borges



UNIVERSIDADE FEDERAL DE UBERLÂNDIA
FACULDADE DE COMPUTAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

Abel Melo Borges

**Aprimorando o atendimento ao cidadão com
RAGLC (Retrieval Augmented Generation in
Logic Context), uma nova abordagem RAG**

Dissertação de mestrado apresentada ao Programa de Pós-graduação da Faculdade de Computação da Universidade Federal de Uberlândia como parte dos requisitos para a obtenção do título de Mestre em Ciência da Computação.

Área de concentração: Ciência da Computação

Orientador: Prof. Dr. Jefferson Rodrigo de Souza

Coorientador: Dr. RaulCezar Maximiano Figueira Alves

Uberlândia

2025

Dados Internacionais de Catalogação na Publicação (CIP)
Sistema de Bibliotecas da UFU, MG, Brasil.

- B732a Borges, Abel Melo, 1984-
2025 Aprimorando o atendimento ao cidadão com RAGLC (*Retrieval Augmented Generation in Logic Context*), uma nova abordagem RAG [recurso eletrônico] / Abel Melo Borges. - 2025.

Orientador: Jefferson Rodrigo de Souza.
Coorientador: Raulcézar Maximiano Figueira Alves.
Dissertação (Mestrado) - Universidade Federal de Uberlândia, Programa de Pós-graduação em Ciência da Computação.
Modo de acesso: Internet.
Disponível em: <http://doi.org/10.14393/ufu.di.2025.5207>
Inclui bibliografia.
Inclui ilustrações.

1. Computação. I. Souza, Jefferson Rodrigo de, 1985-, (Orient.). II. Alves, Raulcézar Maximiano Figueira, 1984-, (Coorient.). III. Universidade Federal de Uberlândia. Programa de Pós-graduação em Ciência da Computação. IV. Título.

CDU: 681.3

André Carlos Francisco
Bibliotecário-Documentalista - CRB-6/3408



UNIVERSIDADE FEDERAL DE UBERLÂNDIA

Coordenação do Programa de Pós-Graduação em Computação
Av. João Naves de Ávila, 2121, Bloco 1A, Sala 243 - Bairro Santa Mônica, Uberlândia-MG,
CEP 38400-902
Telefone: (34) 3239-4470 - www.ppgco.facom.ufu.br - cpgfacom@ufu.br



ATA DE DEFESA - PÓS-GRADUAÇÃO

Programa de Pós-Graduação em:	Ciência da Computação			
Defesa de:	Dissertação, 14/2025, PPGCO			
Data:	31 de Julho de 2025	Hora de início:	13:30	Hora de encerramento:
Matrícula do Discente:	12312CCP003			
Nome do Discente:	Abel Melo Borges			
Título do Trabalho:	Aprimorando o atendimento ao cidadão com RAGLC (Retrieval Augmented Generation in Logic Context), uma nova abordagem RAG			
Área de concentração:	Ciência da Computação			
Linha de pesquisa:	Inteligência Artificial			
Projeto de Pesquisa de vinculação:	-----			

Reuniu-se por videoconferência, a Banca Examinadora, designada pelo Colegiado do Programa de Pós-graduação em Ciência da Computação, assim composta: Professores Doutores: RaulCezar Maximiano Figueira Alves - FACOM/UFU (Coorientador), Leandro Nogueira Couto - FACOM/UFU, Thomas Kauê Dal'Maso Peron - ICMC/USP e Jefferson Rodrigo de Souza - FACOM/UFU, orientador do candidato.

Os examinadores participaram desde as seguintes localidades: Thomas Kauê Dal'Maso Peron - São Carlos/SP. Os outros membros da banca e o aluno participaram da cidade de Uberlândia.

Iniciando os trabalhos o presidente da mesa, Prof. Dr. Jefferson Rodrigo de Souza , apresentou a Comissão Examinadora e o candidato, agradeceu a presença do público, e concedeu ao Discente a palavra para a exposição do seu trabalho. A duração da apresentação da Discente e o tempo de arguição e resposta foram conforme as normas do Programa.

A seguir o senhor presidente concedeu a palavra, pela ordem sucessivamente, aos examinadores, que passaram a arguir ao candidato. Ultimada a arguição, que se desenvolveu dentro dos termos regimentais, a Banca, em sessão secreta, atribuiu o resultado final, considerando o candidato:

Aprovado

Esta defesa faz parte dos requisitos necessários à obtenção do título de Mestre.

O competente diploma será expedido após cumprimento dos demais requisitos, conforme as normas do Programa, a legislação pertinente e a regulamentação interna da UFU.

Nada mais havendo a tratar foram encerrados os trabalhos. Foi lavrada a presente ata que após lida e achada conforme foi assinada pela Banca Examinadora.



Documento assinado eletronicamente por **Jefferson Rodrigo de Souza**, **Professor(a) do Magistério Superior**, em 04/08/2025, às 15:32, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do [Decreto nº 8.539, de 8 de outubro de 2015](#).



Documento assinado eletronicamente por **Thomas Kauê Dal'Maso Peron**, **Usuário Externo**, em 05/08/2025, às 12:27, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do [Decreto nº 8.539, de 8 de outubro de 2015](#).



Documento assinado eletronicamente por **Raulcézar Maximiano Figueira Alves**, **Analista de Tecnologia da Informação**, em 08/08/2025, às 09:58, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do [Decreto nº 8.539, de 8 de outubro de 2015](#).



Documento assinado eletronicamente por **Leandro Nogueira Couto**, **Professor(a) do Magistério Superior**, em 13/08/2025, às 14:52, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do [Decreto nº 8.539, de 8 de outubro de 2015](#).



A autenticidade deste documento pode ser conferida no site https://www.sei.ufu.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **6488728** e o código CRC **987EB0E4**.

Referência: Processo nº 23117.046332/2025-95

SEI nº 6488728

Dedico este trabalho à minha família, que sempre esteve ao meu lado em todos os momentos, compartilhando amor, apoio e inspiração. Sem vocês, nada disso seria possível.

Agradecimentos

Antes de tudo, agradeço a Deus, que me deu força, perseverança e sabedoria para enfrentar os desafios e superar as adversidades ao longo desta jornada.

Ao meu filho Pedro, minha maior inspiração e a melhor expressão de mim mesmo, que dá sentido ao meu trabalho e esforços.

Aos meus pais, que sempre foram meu esteio e a base sólida sobre a qual construí minha vida, oferecendo amor, apoio e valores inestimáveis.

À minha irmã, sobrinhas e ao meu cunhado, por seu carinho, companheirismo e por serem parte essencial da minha história. Vocês representam amor, apoio e alegria em minha vida, e sou profundamente grato a cada um de vocês.

À Luana e Ana Júlia, pela compreensão, amor e apoio que foram essenciais em todo o caminho.

Agradeço também ao Dr. Paulo, ao Dr. Humberto da Secretaria Municipal de Administração e aos meus colegas e professores do programa de Pós-Graduação em Ciência da Computação, que me acompanharam nesta caminhada acadêmica. O suporte, as discussões enriquecedoras e a troca de conhecimentos foram cruciais para a realização deste trabalho.

“A fé é dar o primeiro passo, mesmo quando você não vê toda a escada.”

(Martin Luther King Jr.)

Resumo

Este trabalho desenvolve e avalia um assistente inteligente, fundamentado em Geração Aumentada por Recuperação (RAG) e Grandes Modelos de Linguagem (LLMs), para automatizar e aprimorar o atendimento aos candidatos de concursos públicos, bem como os novos admitidos na Prefeitura Municipal de Uberlândia. Ele aborda conceitos baseados em Processamento de Linguagem Natural (PLN) e Inteligência Artificial de baixo custo para o órgão. A crescente demanda e oferta de serviços pela Prefeitura, e o número cada vez maior de *turnover* entre os funcionários, faz com que o número de contratações municipais e o número de servidores seja cada vez maior. Esse movimento gera a sobrecarga para os servidores das áreas de recrutamento de pessoal, que mesmo com muito esforço geram insatisfação no atendimento de candidatos e ingressantes no corpo de servidores do município. Todo esse contexto motivou a criação de uma solução que eleva a eficiência, precisão e celeridade nesse serviço público. A metodologia emprega uma arquitetura RAG avançada, que se distingue pela proposição de uma técnica inovadora de segmentação textual baseada em coerência semântica, superando os métodos convencionais de divisão por tamanho fixo. Foi realizada uma análise, comparando múltiplas estratégias de recuperação, incluindo o enriquecimento das perguntas dos usuários promovida por LLM, e avaliando o impacto de diferentes modelos de embeddings na qualidade das respostas. Os resultados incluem a avaliação de um modelo eficiente e capaz de processar demandas de mais de 80.000 candidatos que prestaram os concursos e processos seletivos da Prefeitura Municipal de Uberlândia nos últimos 4 anos, demonstrando como a automação unida à Inteligência Artificial pode melhorar a qualidade do atendimento público. Conclui-se que o uso estratégico de tecnologias de IA, sistemas RAG com recuperação de contexto otimizada, pode não apenas otimizar processos administrativos, mas estabelecer um novo paradigma de excelência, transparência e acessibilidade no serviço público.

Palavras-chave: Processamento de linguagem natural. Chatbots. LLM. Administração Pública. Automação. Atendimento ao cidadão.

Abstract

This work develops and evaluates an intelligent conversational assistant, based on Retrieval Augmented Generation (RAG) and Large Language Models (LLMs), to automate and improve service to candidates for public examinations and selection processes, as well as new hires at the Uberlândia City Hall. It addresses concepts based on Natural Language Processing (NLP) and low-cost Artificial Intelligence for the Agency. The growing demand and supply of services by the City Hall, and the increasing number of *turnover* among employees, means that the number of municipal hires and the number of civil servants is increasing. This movement directly overloads the civil servants in the personnel recruitment areas, which, despite great effort, generates dissatisfaction in the service provided by many candidates and newcomers to the municipal staff. This entire context motivated the creation of a solution that increases the efficiency, precision and speed of this public service. The methodology employs an advanced RAG architecture, which stands out for proposing an innovative text segmentation technique based on semantic coherence, outperforming conventional fixed-length division methods. A rigorous empirical analysis was performed, comparing multiple retrieval strategies, including the enrichment of user questions promoted by LLM, and evaluating the impact of different embedding models on the quality of the answers. The expected results include the evaluation of an efficient model capable of processing demands from more than 80,000 candidates who took part in the public examinations and selection processes of the City Hall of Uberlândia in the last 4 years, demonstrating how automation combined with Artificial Intelligence can improve the quality of public service. It is concluded that the strategic use of AI technologies, specifically RAG systems with optimized context recovery, can not only improve administrative processes, but also establish a new paradigm of excellence, transparency, and accessibility in public service.

Keywords: Natural language processing. Chatbots. Public administration. Automation. Citizen service..

Lista de ilustrações

Figura 1 – Fluxo principal que representa todo o modelo. (Elaborado pelo autor)	63
Figura 2 – Representação de funcionalidade do modelo RAGLC. (Elaborado pelo autor)	65
Figura 3 – Fluxograma do Retrieval Augmented Generation / Geração Aumentada por Recuperação (RAG) Tradicional (Naive RAG) com divisão de chunks por tamanho fixo. (Elaborado pelo autor)	74
Figura 4 – Fluxograma do RAG por Bloco, com divisão baseada em unidades semânticas do documento. (Elaborado pelo autor)	75
Figura 5 – Fluxograma do RAG Semântico (RAGLC) proposto, que adiciona uma etapa de otimização contextual da pergunta via LLM. (Elaborado pelo autor)	75
Figura 6 – Comparando as seis estratégias de RAG. (Elaborado pelo Autor)	79
Figura 7 – Resposta bem sucedida do modelo. Contexto recuperado corretamente. (Elaborado pelo Autor)	82
Figura 8 – Resposta do Agente que não conseguiu recuperar o contexto correto. (Elaborado pelo Autor)	83

Lista de tabelas

Tabela 1 – Comparativo entre Técnicas de Segmentação. (Elaborada pelo autor)	39
Tabela 2 – Comparativo de modelos de embeddings. (Elaborado pelo Autor)	73
Tabela 3 – Lista de Perguntas sobre o Estatuto dos Servidores. (Elaborado pelo Autor)	76
Tabela 4 – Resultado por blocos de perguntas de modelos RAG	76
Tabela 5 – Resultados de pontuação (0–10) para cada uma das 20 perguntas.	77

Lista de siglas

API Application Programming Interface

AG Algorítmo Genético

CPF Cadastro de Pessoa Física

DL Deep Learning

HTTP Hypertext Transfer Protocol

IA Inteligência Artificial

LSTM Long Short-Term Memory

LLM Large Language Model / Grandes Modelos de Linguagem

LGPD Lei Geral de Proteção de Dados

ML Machine Learning / Aprendizado de Máquina

NLP Natural Language Processing

PLN Processamento de Linguagem Natural

PDF Portable Document Format

RAG Retrieval Augmented Generation / Geração Aumentada por Recuperação

RNN Recurrent Neural Network / Redes Neurais Recorrentes

RLHF Aprendizado por Reforço com Feedback Humano

RAGLC Retrieval Augmented Generation in Logic Context

SQL Structured Query Language

Sumário

1	INTRODUÇÃO	25
1.1	Motivação	26
1.2	Objetivos e Desafios da Pesquisa	27
1.3	Hipótese	28
1.4	Contribuições	28
1.5	Organização da Dissertação	29
2	FUNDAMENTAÇÃO TEÓRICA	31
2.1	Conceitos Fundamentais de AI	31
2.1.1	Definições e Taxonomias da AI	32
2.1.2	Arquiteturas de Redes Neurais e DL	33
2.1.3	Embeddings Distribuídos: A Revolução Semântica	34
2.2	Grandes Modelos de Linguagem (LLMs)	34
2.2.1	Pré-processamento Interno da API DeepSeek	35
2.2.2	Fundamentos Teóricos e Arquiteturais do RAG	35
2.2.3	Evolução e Estado da Arte em RAG (2022-2025)	37
2.3	Tecnologias e Ferramentas	40
2.3.1	Linguagens de Programação e Ambientes de Desenvolvimento	41
2.3.2	Frameworks de DL e Computação Neural	41
2.3.3	Plataformas de Modelos e Recursos Pré-Treinados	41
2.3.4	Frameworks de Orquestração RAG	42
2.3.5	Bancos de Dados Vetoriais e Sistemas de Busca	42
2.3.6	Modelos de Embedding Especializados	42
2.3.7	Ferramenta de Interface	43
2.3.8	Plataformas de Automação e Integração: N8N	43
2.3.9	APIs de Comunicação Avançada: Evolution API	43
2.4	Configuração Utilizada (<i>setup</i>)	43
2.5	Trabalhos Relacionados e Contexto Científico	44

2.5.1	Aplicações de LLMs no Setor Público	45
2.5.2	Governo Algorítmico e IA Pública	46
2.5.3	Retrieval-Augmented Generation for Large Language Models: A Survey	46
2.5.4	Posicionamento Científico e Contribuições Distintivas	48
2.6	Considerações Finais da Fundamentação Teórica	48
3	METODOLOGIA PROPOSTA E DESENVOLVIMENTO	51
3.1	Visão Geral do Pipeline Experimental	51
3.2	Definição do Problema e Coleta de Dados	52
3.3	Validação Preliminar: Canal de Comunicação	52
3.4	Análise da melhor estratégia para resolução do problema específico	53
3.5	Avaliação de Modelos de Embeddings	54
3.6	Algoritmo de Segmentação por Similaridade Semântica	55
3.7	Dimensões de Avaliação da Proposta	58
4	RAG HÍBRIDO PARA INTEGRAÇÃO DE DADOS NO ATENDIMENTO PÚBLICO	61
4.1	Introdução	61
4.2	Fundamentos Teóricos: Consultando Dados Estruturados com Linguagem Natural	62
4.3	Metodologia Proposta: Arquitetura HybridRAG com N8N	63
4.3.1	Arquitetura Geral do Sistema	63
4.3.2	Implementação do Workflow N8N	63
4.3.3	Integração com Google Drive	64
4.4	Implementação Técnica Detalhada	65
4.4.1	Configuração da Evolution API	65
4.4.2	Algoritmo RAGLC	65
4.4.3	Lógica de Roteamento Inteligente	66
4.5	Fluxo Operacional Detalhado	66
4.5.1	Cenário 1: Consulta Apenas a Dados Estruturados	66
4.5.2	Cenário 2: Consulta Apenas a Documentos Não Estruturados	66
4.5.3	Cenário 3: Consulta Híbrida	67
4.6	Vantagens da Arquitetura N8N	67
4.7	Impacto Potencial e Desafios	67
4.8	Integração com a Dissertação	68
4.9	FUNCIONAMENTO DO RAG HÍBRIDO	69
4.10	Conclusão	69

5	EXPERIMENTOS E ANÁLISE DOS RESULTADOS	71
5.1	Método para a Avaliação	71
5.1.1	Avaliação dos Canais de Comunicação	71
5.1.2	Avaliação de Modelos de Embeddings e bancos de dados vetoriais	71
5.1.3	Avaliação da Segmentação Semântica	72
5.1.4	Comparação com Respostas corretas ou Oficiais	72
5.2	Experimentos	72
5.2.1	Validação dos Canais de Comunicação	72
5.2.2	Avaliação de Modelos de Embeddings	73
5.2.3	Algoritmo de Segmentação por Contexto Semântico	74
5.2.4	Avaliação Geral	83
6	CONCLUSÃO	85
6.1	Principais Contribuições	85
6.2	Trabalhos Futuros	86
6.3	Contribuições em Produção Bibliográfica	87
REFERÊNCIAS		89
APÊNDICES		93
APÊNDICE A	–	INTERAÇÃO DO CHATBOT COM O USUÁRIO 95
APÊNDICE B	–	PERGUNTAS E RESPOSTAS CEDIDAS PELA BANCA ORGANIZADORA DE CONCURSOS E PROCESSOS SELETIVOS 101
APÊNDICE C	–	RESULTADOS DO ÚLTIMO EXPERIMENTO POR MODELO EM CADA MÉTRICA 109

CAPÍTULO 1

Introdução

A transformação digital tem revolucionado setores públicos e privados, impulsionada pela necessidade de eficiência e acessibilidade. No Brasil, iniciativas como o portal "gov.br" exemplificam esse avanço, centralizando serviços governamentais com economia anual estimada em R\$ 100 milhões e investimentos de R\$ 43 milhões até 2020 (BRASIL, 2021a); (BRASIL, 2021b). Essa digitalização visa transparência e alinhamento às demandas da sociedade moderna, refletindo um esforço contínuo por gestão pública eficaz.

A Prefeitura Municipal de Uberlândia destaca-se como referência em inovação tecnológica. Premiada em 2021 e 2022 como uma das "Cidades com Serviços Inteligentes" (UBERLÂNDIA, 2021); (UBERLÂNDIA, 2022c), o município adotou práticas pioneiras, como reuniões e homologações de licitações no metaverso (UBERLÂNDIA, 2022a); (UBERLÂNDIA, 2022b). Nesse cenário, a Prefeitura Municipal de Uberlândia tem se destacado como um polo de inovação tecnológica, sendo premiada como uma das "Cidades com Serviços Inteligentes" em 2021 e 2022. A adoção de práticas pioneiras, como homologações de licitações no metaverso, sinaliza uma cultura organizacional propensa à experimentação e aos desafios motivados pela tecnológica.

Apesar dos avanços estratégicos, desafios operacionais persistem. O Núcleo de Recrutamento da Secretaria Municipal de Administração, responsável pela gestão de recrutamento a partir de concursos e processos seletivos, enfrenta um volume expressivo e flutuante de demandas. O setor lida com cerca de 200 atendimentos diárias entre telefônicos e por mensagens, sendo este último mais de 80% da demanda. Esse fluxo se intensifica em períodos críticos, como o início do ano letivo e volta das férias escolares. Com uma equipe de apenas oito funcionários para gerenciar anualmente cerca de aproximadamente 12.000 candidatos e 15.000 servidores, a sobrecarga é inevitável.

Esta sobrecarga resulta em consequências mensuráveis e impactantes:

- ❑ Alto Tempo Médio de Atendimento (TMA): Respostas a e-mails e mensagens podem levar horas ou dias, gerando ansiedade e insatisfação nos candidatos.
- ❑ Baixa Taxa de Resolução no Primeiro Contato: A complexidade e a repetitividade

das dúvidas sobre editais consomem um tempo desproporcional dos servidores, que poderiam se dedicar a tarefas de maior valor agregado.

- Insatisfação dos Usuários e Servidores: O ciclo de espera e a sobrecarga de trabalho criam um ambiente de frustração mútua. Em momentos de alto pico é notório o crescente afastamento desses servidores para tratamento de saúde e o declínio motivacional da equipe.

Para mitigar esses problemas, este trabalho propõe o desenvolvimento e a avaliação de um chatbot especializado, funcionando dentro do aplicativo de mensagens WhatsApp, fundamentado em Natural Language Processing (NLP) e Large Language Model / Grandes Modelos de Linguagem (LLM). Esses sistemas, capazes de simular interações humanas por texto, têm se destacado em áreas como atendimento ao cliente e saúde (CHANDEL; SHUKLA; GUPTA, 2018). Sistemas conversacionais baseados em Inteligência Artificial (IA) já demonstraram potencial para otimizar respostas, reduzir custos operacionais e melhorar a satisfação dos usuários em diversos setores (BRASIL, 2024).

Este trabalho visa desenvolver e avaliar um Assistente Conversacional multitarefas para o Núcleo de Recrutamento de Uberlândia, analisando sua viabilidade em fornecer respostas precisas a demandas recorrentes, como dúvidas sobre editais de concursos, pesquisa em banco de dados interno para verificar sobre o status de uma convocação ou responder perguntas sobre planos de carreiras, quantidade de vagas, benefícios ofertados aos servidores municipais ou referentes ao Estatuto do Servidor. A ferramenta utilizará um LLM treinados com dados contextuais, incluindo textos segmentados por critérios semânticos (avaliando a estrutura dos documentos importados) em vez da divisão convencional por quantidade de palavras. Essa abordagem inovadora demonstrou superioridade em testes preliminares com modelos de embeddings (BRASIL, 2025) específicos. Adicionalmente, o estudo comparará respostas do chatbot com respostas humanas oficiais, validando sua eficácia por meio de métricas de precisão e relevância.

A pesquisa contribui para o debate sobre a aplicação de IA no setor público, destacando seu potencial para replicar capacidades humanas (como raciocínio lógico e interpretação) em tarefas operacionais, que exigem grande esforço repetitivo e massante do ser humano. Ao integrar tecnologias emergentes a administração pública pode alcançar maior eficiência e transparência, alinhando-se às expectativas.

1.1 Motivação

A inovação no setor público é historicamente reconhecida como motor de eficiência e modernização (SCHUMPETER, 1934; OSBORNE; BROWN, 2005). Contudo, ainda existem lacunas significativas na adoção de tecnologias emergentes voltadas especificamente para o atendimento ao cidadão. Embora haja estudos sobre informatização de

serviços (PAIVA; FERNANDES; ALTURAS, 2022), poucas pesquisas investigam o uso de IA e NLP para automatizar, de forma inteligente, a interpretação de documentos legais e administrativos ou visam desenvolver uma solução para algum problema da Administração Pública.

Nesse cenário, a Prefeitura Municipal de Uberlândia se apresenta como um campo propício para experimentação tecnológica. Este estudo concentra-se na implementação de um chatbot que responde automaticamente a demandas de concursos públicos e processos seletivos, utilizando técnicas de IA para processar linguagem natural e integrar dados estruturados e não estruturados.

A solução visa reduzir a carga de trabalho dos servidores, aumentar a agilidade do atendimento e contribuir para a economia de recursos públicos, estimada em até meio milhão de reais anuais entre custos operacionais e de pessoal.

1.2 Objetivos e Desafios da Pesquisa

O objetivo geral deste estudo é desenvolver, avaliar e validar uma arquitetura de RAG para um chatbot (dentro do WhatsApp) que interprete e responda, com alta fidelidade e de forma automatizada, a consultas de candidatos ou ingressantes no serviço público municipal referentes aos processos e as relações de trabalho. Além disso, o trabalho busca avaliar a viabilidade da implantação de tecnologias de código aberto ou de baixo custo em ambientes públicos, considerando seus impactos operacionais, financeiros e sua capacidade de escalabilidade.

O estudo tem como foco comparar técnicas consolidadas da literatura e do mercado para a transformação de texto e tokens em embeddings, utilizando um modelo baseado em segmentação por contextualização ao invés das técnicas de divisão por número de caracteres e *overlap*. Em outro cenário, é avaliada a qualidade da contextualização em modelos RAG, comparando técnicas tradicionais com uma abordagem inovadora de contextualização por tópicos, em contrapartida ao modelo tradicional de divisão por número de caracteres e *overlap*.

Os objetivos específicos foram definidos para abordar sistematicamente este desafio:

- ❑ Desenvolver e avaliar uma técnica de segmentação semântica para documentos jurídicos/administrativos, com o objetivo de melhorar a qualidade do contexto recuperado quando comparada aos métodos tradicionais.
- ❑ Comparar múltiplos modelos de embeddings para identificar o mais eficaz na representação de textos normativos e editais em português brasileiro, especificamente da Prefeitura Municipal de Uberlândia.
- ❑ Analisar o impacto do enriquecimento de consultas (*query enrichment*) na precisão e relevância das respostas geradas pelo sistema RAG.

- Validar a usabilidade e a eficácia do modelo através de uma comparação direta com as respostas oficiais humanas, mensurando a completude e a personalização.
- Propor uma arquitetura de RAG Híbrido como um avanço conceitual, capaz de integrar dados não estruturados (editais ou leis da Prefeitura Municipal de Uberlândia) com dados estruturados (planilhas ou bancos de dados de candidatos contendo o andamento de suas convocações) para viabilizar um atendimento completo .

O desenvolvimento desta pesquisa envolve desafios, como a adaptação de modelos de IA às especificidades da linguagem administrativa e jurídica dos editais de processos seletivos, a integração eficiente da tecnologia aos sistemas existentes e a superação de barreiras culturais e organizacionais relacionadas à adoção de novas tecnologias na administração pública, mesmo em um ambiente mais propício que em outros órgãos. Além disso, aspectos críticos relacionados à segurança da informação, privacidade dos dados dos usuários e transparência nos processos automatizados foram cuidadosamente considerados, assegurando que a implementação da solução tecnológica esteja em conformidade com as melhores práticas de governança e ética na administração pública.

1.3 Hipótese

Diante dos desafios enfrentados na otimização do atendimento ao público em órgãos públicos, especialmente no contexto da Prefeitura Municipal de Uberlândia, este trabalho propõe testar a seguinte hipótese:

A aplicação de um sistema RAG, otimizado com segmentação semântica contextual e enriquecimento da consulta, melhora significativamente a qualidade (medida por métricas de precisão e relevância como BERTSCORE), a celeridade e a completude do atendimento a candidatos de concursos públicos, quando comparado aos métodos tradicionais (atendimento humano via e-mail) e a baselines de RAG convencionais.

1.4 Contribuições

As contribuições desta pesquisa foram:

1. Proposição e Validação de uma Metodologia de Segmentação Semântica: Foi demonstrado que a segmentação baseada em coerência semântica supera as abordagens tradicionais de divisão por tamanho fixo em documentos com estrutura lógica complexa, como editais e leis.
2. Análise Comparativa de Estratégias de RAG: Foi fornecida uma análise detalhada sobre o desempenho combinado de diferentes técnicas de segmentação, modelos de

embeddings e estratégias de enriquecimento de consulta, oferecendo um guia prático para a otimização de sistemas RAG no domínio público-administrativo.

3. Proposta de uma Arquitetura de RAG Híbrido para uma estrutura de chatbot: Foi apresentado um modelo conceitual para a integração de fontes de dados estruturados e não estruturados em sistemas conversacionais, abordando uma limitação crítica dos sistemas RAG atuais.
4. Implementado de um Chatbot de Baixo Custo, dentro do aplicativo de mensagens WhatsApp: Foi desenvolvido e avaliado um chatbot funcional, cuja arquitetura e código-fonte são disponibilizados para replicação por outros órgãos públicos, promovendo a democratização da tecnologia, através da utilização de ferramentas, frameworks *opensource*.
5. Melhora na Prestação de Serviço Público: A solução proposta tem o potencial de reduzir drasticamente o tempo de resposta, aumentar a disponibilidade do serviço (24/7) e melhorar a satisfação dos candidatos, gerando economia aos cofres públicos e liberando servidores para atividades estratégicas.
6. Contribuição para a modernização da administração pública, oferecendo uma solução para outros órgãos interessados na adoção de tecnologias baseadas em IA para otimização de processos e melhoria da experiência do usuário.

1.5 Organização da Dissertação

Esta dissertação está organizada em seis capítulos que abordam os aspectos da pesquisa.

No **Capítulo 1**, apresenta-se a introdução ao tema, contextualizando o escopo da pesquisa, expondo a motivação, os objetivos, a hipótese, as contribuições e delineando a estrutura geral do trabalho.

O **Capítulo 2** aprofunda a fundamentação teórica, discutindo conceitos relacionados à IA, NLP, modelos de LLM, algoritmos genéticos, tokenização, e representação vetorial de textos, além de apresentar uma revisão extensiva do estado da arte em RAG e tecnologias associadas.

O **Capítulo 3** detalha a metodologia proposta e desenvolvimento, apresentando os métodos utilizados para validação dos canais de comunicação, avaliação de usabilidade do chatbot, otimização da recuperação de contexto com K-means e algoritmos genéticos, algoritmo de segmentação por contexto semântico, avaliação de modelos de embeddings e comparação entre o chatbot e as respostas oficiais.

O **Capítulo 4** introduz uma arquitetura de RAG Híbrido para integração de dados estruturados e não estruturados no atendimento público. Este capítulo explora como

combinar consultas a documentos textuais (como editais) com dados estruturados (como planilhas de candidatos), apresentando uma metodologia baseada em agentes inteligentes que direcionam consultas para ferramentas especializadas, aumentando o escopo e a qualidade das respostas fornecidas aos usuários.

O **Capítulo 5** é dedicado aos experimentos e à análise dos resultados, apresentando os testes realizados para validar cada aspecto da solução proposta. São analisados os resultados da validação dos canais de comunicação, usabilidade do chatbot, otimização da recuperação de contexto, algoritmo de segmentação semântica, avaliação comparativa de modelos de embeddings e comparação entre as respostas do chatbot e as respostas oficiais, fornecendo evidências empíricas para validação da hipótese central da pesquisa.

Por fim, o **Capítulo 6** apresenta a conclusão do trabalho, sintetizando as principais contribuições da pesquisa, destacando como os objetivos específicos foram alcançados, discutindo as limitações do estudo e sugerindo direções para trabalhos futuros, além de listar as produções bibliográficas resultantes da pesquisa.

CAPÍTULO 2

Fundamentação Teórica

Este capítulo estabelece as bases teóricas que sustentam a pesquisa, mergulhando nos conceitos fundamentais de IA, NLP, LLMs e, com especial atenção, na arquitetura de RAG (LEWIS et al., 2020b). Além de cobrir os fundamentos essenciais, esta seção apresenta uma revisão crítica e abrangente do estado da arte, focando nos avanços mais recentes (2022-2025) que moldam o campo e contextualizam as contribuições.

A fundamentação teórica aborda os principais conceitos e técnicas em IA e NLP (RUSSELL; NORVIG, 2020a), discutindo os aspectos teóricos e práticos da geração automática de texto, Machine Learning / Aprendizado de Máquina (ML) e Deep Learning (DL), bem como as metodologias que permitem representar e manipular dados textuais de forma eficiente e precisa. O capítulo culmina na discussão dos grandes modelos de linguagem e suas aplicações práticas, incluindo as tecnologias e ferramentas que viabilizam a implementação de sistemas inteligentes modernos.

Ademais, são apresentados trabalhos relacionados e suas abrangências no atendimento de demandas que envolvem o setor público e privado, demonstrando a relevância e aplicabilidade dos conceitos abordados em contextos reais e desafiadores da sociedade contemporânea.

2.1 Conceitos Fundamentais de AI

A IA representa um dos campos dinâmicos e transformadores da ciência da computação contemporânea (RUSSELL; NORVIG, 2020a), constituindo-se como uma área interdisciplinar que combina elementos da matemática, estatística, ciência da computação, neurociência e filosofia. Desde sua concepção formal na década de 1950, durante a histórica Conferência de Dartmouth (MCCARTHY et al., 2006), a IA tem evoluído de conceitos teóricos abstratos para aplicações práticas que permeiam diversos aspectos da sociedade moderna, desde sistemas de recomendação em plataformas digitais até assistentes virtuais inteligentes e sistemas de diagnóstico.

2.1.1 Definições e Taxonomias da AI

A definição de IA tem sido objeto de intenso debate acadêmico desde os primórdios do campo (RUSSELL; NORVIG, 2020a), refletindo a complexidade inerente ao conceito de "inteligência" e sua implementação. Russell e Norvig (RUSSELL; NORVIG, 2020b) propõem uma taxonomia quadripartite aceita na comunidade científica, que categoriza as definições de IA em quatro dimensões fundamentais: sistemas que pensam como humanos, sistemas que agem como humanos, sistemas que pensam racionalmente e sistemas que agem racionalmente. Esta classificação oferece uma estrutura conceitual robusta e abrangente para compreender as diferentes abordagens, metodologias e objetivos dentro do vasto campo da IA.

A primeira categoria, **sistemas que pensam como humanos**, baseia-se na premissa fundamental de que a IA deve replicar os processos cognitivos humanos em sua essência e funcionamento. Esta abordagem, influenciada pela ciência cognitiva e pela psicologia experimental, busca modelar os mecanismos mentais subjacentes ao raciocínio, aprendizagem, memória e tomada de decisão humanos. Embora conceitualmente atrativa e filosoficamente interessante, esta perspectiva enfrenta limitações devido à complexidade extraordinária e à incompletude do conhecimento atual sobre o funcionamento da mente humana, incluindo os processos neurobiológicos subjacentes à cognição.

A segunda categoria, **sistemas que agem como humanos**, concentra-se primordialmente no comportamento observável e mensurável, nos processos internos ou mecanismos cognitivos subjacentes. O famoso Teste de Turing, proposto por Alan Turing em 1950 (TURING, 1950), exemplifica esta abordagem behaviorista, propondo que um sistema computacional pode ser considerado inteligente se conseguir enganar um interrogador humano experiente em uma conversa textual estruturada, demonstrando comportamento indistinguível do humano. Esta perspectiva behaviorista tem influenciado o desenvolvimento contemporâneo de interfaces conversacionais avançadas, chatbots sofisticados e assistentes virtuais inteligentes.

A terceira categoria, **sistemas que pensam racionalmente**, fundamenta-se solidamente na tradição lógica formal e filosófica milenar, buscando criar sistemas computacionais que raciocinem de acordo com princípios lógicos formais rigorosos e demonstráveis. Esta abordagem, enfrenta desafios práticos substanciais relacionados à formalização adequada do conhecimento do mundo real em estruturas lógicas manipuláveis, bem como à tratabilidade de problemas complexos que envolvem raciocínio sobre domínios extensos.

A quarta categoria, **sistemas que agem racionalmente**, representa a abordagem predominante e mais pragmática na IA contemporânea, definindo um agente racional como aquele que age consistentemente de forma a maximizar seu desempenho esperado, dado seu conhecimento disponível, capacidades computacionais e restrições ambientais. Esta definição pragmática e orientada a resultados tem permitido o desenvolvimento bem-sucedido de sistemas de IA altamente eficazes em domínios específicos e aplicações práticas

diversas.

2.1.2 Arquiteturas de Redes Neurais e DL

O ressurgimento espetacular das redes neurais artificiais, impulsionado pelos avanços revolucionários em DL e pela disponibilidade de poder computacional massivo, tem revolucionado o campo da IA e criado possibilidades anteriormente para aplicações em diversos domínios. As arquiteturas de DL, caracterizadas por múltiplas camadas de processamento hierárquico e representações distribuídas complexas, têm demonstrado capacidades extraordinárias em tarefas desafiadoras como reconhecimento de padrões visuais, processamento de linguagem natural, geração de conteúdo criativo e modelagem de sistemas dinâmicos complexos.

As **Recurrent Neural Network / Redes Neurais Recorrentes (RNN)**s e suas variantes mais sofisticadas, incluindo Long Short-Term Memory (LSTM) (HOCHREITER; SCHMIDHUBER, 1997) e Gated Recurrent Units (GRUs) (CHO et al., 2014), são especializadas no processamento eficiente de sequências temporais e dados com dependências de longo prazo. Estas arquiteturas incorporam conexões recorrentes que permitem a manutenção de informações ao longo do tempo, possibilitando a modelagem de padrões temporais e dependências contextuais. As LSTMs, em particular, introduziram mecanismos sofisticados de gates (portões) que controlam o fluxo de informações, resolvendo efetivamente o problema do desvanecimento do gradiente que limitava as RNNs tradicionais. Estas arquiteturas têm sido fundamentais para aplicações em processamento de linguagem natural, reconhecimento de fala, análise de séries temporais e modelagem de sequências biológicas.

Os **Transformers**, introduzidos por Vaswani et al. (VASWANI et al., 2017) no trabalho seminal "Attention Is All You Need", representam uma revolução arquitetural que tem dominado completamente o processamento de linguagem natural contemporâneo e expandido sua influência para outros domínios. Baseados em mecanismos de atenção sofisticados, os Transformers eliminam a necessidade de processamento sequencial recorrente, permitindo paralelização eficiente massiva e captura superior de dependências de longo alcance em sequências extensas. O mecanismo de atenção, formalizado matematicamente:

$$\text{Attention}(Q, K, V) = \text{softmax} \left(\frac{QK^T}{\sqrt{d_k}} \right) V$$

onde Q (consultas), K (chaves) e V (valores) são projeções lineares da entrada, e \$d_k\$ é a dimensão dos vetores de chave, permite que o modelo atribua dinamicamente diferentes pesos de importância aos elementos da sequência de entrada, capturando relações complexas e contextos ricos.

Esta arquitetura revolucionária é a base de modelos de linguagem de grande escala como GPT (Generative Pre-trained Transformer) (RADFORD et al., 2018), BERT (Bidi-

rectional Encoder Representations from Transformers) (DEVLIN et al., 2019), T5 (Text-to-Text Transfer Transformer) (RAFFEL et al., 2020) e seus sucessores mais avançados, que têm demonstrado capacidades impressionantes e surpreendentes em tarefas diversas como geração de texto coerente, tradução automática de alta qualidade, resposta a perguntas complexas, sumarização automática e até mesmo programação de código.

2.1.3 Embeddings Distribuídos: A Revolução Semântica

Os métodos de embeddings distribuídos representam um avanço paradigmático fundamental na representação textual, permitindo a codificação de palavras em espaços vetoriais contínuos e densos onde a proximidade geométrica reflete similaridade semântica. Esta abordagem supera as limitações das representações esparsas tradicionais, capturando relações semânticas complexas e permitindo operações aritméticas com significado linguístico.

2.1.3.1 Sentence Transformers: Embeddings de Sentenças

No trabalho é utilizada a estrutura de **Sentence Transformers** (REIMERS; GUREVYCH, 2019a), que representam uma evolução significativa para a criação de embeddings eficazes a nível de sentença e documento. Baseados em arquiteturas Transformer pré-treinadas, esses modelos, exemplificados pelo Sentence-BERT (REIMERS; GUREVYCH, 2019b), utilizam redes siamesas e triplet networks para ajustar modelos como BERT ou RoBERTa especificamente para gerar embeddings de sentenças que capturam semelhanças semânticas de maneira eficaz e eficiente.

Esta abordagem é valiosa para tarefas que requerem comparação entre sentenças completas, como busca semântica, clustering de documentos, detecção de paráfrases e sistemas de recuperação de informações. Os Sentence Transformers têm demonstrado resultados superiores consistentes em benchmarks de similaridade textual e são amplamente utilizados em sistemas de recomendação, assistentes virtuais e aplicações de recuperação de informações.

2.2 Grandes Modelos de Linguagem (LLMs)

Os LLMs representam uma das conquistas mais significativas e transformadoras da IA contemporânea, constituindo sistemas de NLP que utilizam arquiteturas neurais com bilhões ou até trilhões de parâmetros, treinados em corpora textuais massivos que abrangem uma vasta gama de conhecimento humano. Esses modelos extraordinários, exemplificados por sistemas como GPT (Generative Pre-trained Transformer) (RADFORD et al., 2018), BERT (Bidirectional Encoder Representations from Transformers) (DEVLIN et al., 2019), LLaMA (Large Language Model Meta IA) (TOUVRON et al., 2023), T5 (Text-to-Text Transfer Transformer) (RAFFEL et al., 2020) e DeepSeek (TEAM, 2023),

têm demonstrado habilidades surpreendentes e frequentemente emergentes em diversas tarefas linguísticas complexas, desde geração de texto coerente e criativo até tradução automática de alta qualidade, sumarização inteligente, resposta a perguntas complexas, e até mesmo programação de código.

A escalabilidade impressionante desses modelos, combinada com sua capacidade de capturar contextos extensos, padrões linguísticos sutis e conhecimento factual abrangente, os torna ferramentas extraordinariamente poderosas e versáteis tanto para aplicações acadêmicas quanto industriais. Sua capacidade de realizar tarefas de few-shot e zero-shot learning, onde podem executar tarefas novas com poucos ou nenhum exemplo específico de treinamento, representa um avanço paradigmático na direção de uma IA mais geral e adaptável.

2.2.1 Pré-processamento Interno da API DeepSeek

Este segmento descreve o fluxo interno de uma requisição à API DeepSeek (Utilizada no trabalho), conforme a documentação oficial, a qual usa como padrão a mesma estrutura de acesso que é utilizada pela OpenAI, com ênfase nas etapas essenciais:

1. **Tokenização e Prompt:** A **tokenização** representa a primeira e uma das mais importantes etapas no pipeline de processamento de linguagem natural, consistindo na segmentação sistemática do texto bruto em unidades menores denominadas tokens. O texto é segmentado em tokens, que podem representar palavras, caracteres ou símbolos. Aproximadamente, um caractere em inglês se converte em 0,3 token; em chinês, 0,6 token (DeepSeek AI, 2024b). O payload inclui controle de ‘max_tokens’ e penalidades de repetição (‘presence_penalty’), e permite o retorno estruturado via ‘response_format = {"type": "json_object"}’ (DeepSeek AI, 2024c).
2. **Batching e Streaming:** A API suporta formatos compatíveis com OpenAI, possibilitando o uso de SDKs. O parâmetro ‘stream’ permite a emissão incremental da resposta via SSE, essencial para interfaces em tempo real (como WhatsApp) (DeepSeek AI, 2024a).
3. **Decoding com Arquitetura Avançada:** O modelo DeepSeek-V3 emprega Mixture-of-Experts (ativando apenas 37B de 671B parâmetros por token) e Multi-head Latent Attention para compressão de KV-cache. Além disso, utiliza multi-token prediction, capaz de gerar múltiplos tokens por passo, melhorando a eficiência (DeepSeek AI, 2024d).

2.2.2 Fundamentos Teóricos e Arquiteturais do RAG

A arquitetura RAG fundamenta-se na premissa de que a qualidade da geração de texto pode ser substancialmente melhorada através do acesso dinâmico a informações

relevantes e atualizadas. O processo canônico de RAG envolve duas etapas principais interconectadas e complementares:

2.2.2.1 Etapa de Recuperação (Retrieval)

Na primeira etapa, dado uma consulta ou pergunta do usuário, um sistema de recuperação especializado (retriever) busca e identifica os trechos de informação mais relevantes e semanticamente relacionados em uma base de conhecimento externa. Esta busca é tipicamente realizada através de técnicas de busca por similaridade vetorial, onde tanto a consulta quanto os documentos da base de conhecimento são representados como vetores densos em um espaço semântico compartilhado.

O processo de recuperação pode ser formalizado como uma função R que mapeia uma consulta q para um conjunto de documentos relevantes:

$$R(q) = \{d_1, d_2, \dots, d_k\} \subset \mathcal{D}$$

onde \mathcal{D} representa a base de conhecimento completa e k é o número de documentos recuperados (tipicamente entre 3 e 10).

A qualidade da recuperação é fundamental para o sucesso do sistema RAG completo, sendo tipicamente avaliada através de métricas como:

- ❑ **Recall@k:** Proporção de documentos relevantes recuperados entre os top-k
- ❑ **Precision@k:** Proporção de documentos recuperados que são realmente relevantes
- ❑ **Mean Reciprocal Rank (MRR):** Média dos inversos das posições do primeiro documento relevante

2.2.2.2 Etapa de Geração (Generation)

Na segunda etapa, a consulta original do usuário e os documentos recuperados são combinados de forma estruturada em um prompt expandido e contextualizado, que é então fornecido ao modelo de linguagem gerador. O LLM utiliza tanto a consulta original quanto o contexto recuperado para gerar uma resposta fundamentada, precisa e contextualmente apropriada.

A probabilidade condicional da resposta gerada pode ser expressa matematicamente como:

$$P(y|x) = \sum_{r \in \mathcal{R}(x)} P(y|x, r) \cdot P(r|x)$$

onde:

- ❑ y é a resposta gerada
- ❑ x é a consulta original

- ❑ $\mathcal{R}(x)$ é o conjunto de documentos recuperados para a consulta x
- ❑ $P(r|x)$ é a probabilidade de relevância do documento r para a consulta x
- ❑ $P(y|x, r)$ é a probabilidade de gerar resposta y dado a consulta x e o documento r

2.2.3 Evolução e Estado da Arte em RAG (2022-2025)

A pesquisa em RAG tem evoluído rapidamente desde sua introdução, passando de implementações relativamente simples e diretas para abordagens modulares, sofisticadas e altamente otimizadas. Esta evolução pode ser sistematicamente categorizada em três estágios principais do pipeline RAG, cada um com inovações específicas e contribuições significativas para o desempenho geral do sistema.

2.2.3.1 Técnicas de Pré-Recuperação: Otimização da Consulta

O desempenho de sistemas RAG é dependente da qualidade e precisão da consulta inicial utilizada para recuperação. Pesquisas recentes têm focado em técnicas sofisticadas para refinar, expandir e otimizar consultas antes da etapa de busca:

Query Transformation e Expansion: Métodos avançados que utilizam LLMs para reescrever ou expandir sistematicamente a consulta original do usuário, melhorando a eficácia da recuperação. Estas técnicas podem envolver:

- ❑ Geração de múltiplas variantes parafráticas da pergunta original
- ❑ Decomposição de perguntas complexas em sub-perguntas mais específicas e focadas
- ❑ Expansão com termos relacionados e sinônimos relevantes
- ❑ Reformulação em diferentes estilos linguísticos ou níveis de especificidade

Hypothetical Document Embeddings (HyDE): Técnica inovadora proposta por Gao et al. (GAO et al., 2023), que utiliza um LLM para gerar um documento hipotético que teoricamente responderia à consulta do usuário. O embedding deste documento hipotético, que é semanticamente mais próximo dos documentos-alvo reais na base de conhecimento, é então utilizado para a busca vetorial, em vez do embedding da consulta curta original. Esta abordagem tem demonstrado melhorias consistentes na qualidade da recuperação, especialmente para consultas complexas ou ambíguas.

Query Routing e Classification: Sistemas que classificam automaticamente consultas em diferentes categorias ou domínios, direcionando-as para bases de conhecimento especializadas ou estratégias de recuperação otimizadas para cada tipo específico de pergunta.

2.2.3.2 Otimização do Processo de Recuperação

A etapa central de recuperação também tem sido alvo de inovações substanciais e melhorias significativas:

Chunking Semântico Avançado: A segmentação tradicional baseada em tamanho fixo tem sido progressivamente substituída por técnicas mais sofisticadas que consideram coerência semântica, estrutura discursiva e limites conceituais naturais. Abordagens incluem:

- ❑ Segmentação baseada em similaridade de embeddings de sentenças consecutivas
- ❑ Utilização de modelos de segmentação de tópicos como TextTiling e C99
- ❑ Segmentação estrutural baseada em elementos como parágrafos, seções e capítulos
- ❑ Técnicas híbridas que combinam múltiplos critérios de segmentação

Busca Híbrida (Hybrid Search): Combinação sinérgica de busca vetorial semântica (dense retrieval) com busca lexical tradicional baseada em palavras-chave (sparse retrieval, como BM25). Esta abordagem híbrida garante que tanto o significado semântico quanto os termos exatos sejam adequadamente considerados, melhorando a robustez e cobertura da recuperação. A combinação típica utiliza:

$$\text{score}_{hybrid} = \alpha \cdot \text{score}_{dense} + (1 - \alpha) \cdot \text{score}_{sparse}$$

onde α é um parâmetro de平衡amento aprendido ou ajustado empiricamente.

Multi-Vector Retrieval: Técnicas que representam documentos através de múltiplos vetores (por exemplo, um para cada sentença ou parágrafo), permitindo recuperação mais granular e precisa de informações específicas dentro de documentos longos.

2.2.3.3 Técnicas de Pós-Recuperação: Re-ranking e Refinamento

Após a recuperação inicial, os documentos candidatos podem ser processados através de técnicas sofisticadas para melhorar sua relevância, precisão e eficiência:

Re-ranking com Cross-Encoders: Utilização de modelos mais sofisticados e intensivos (cross-encoders) para reordenar a lista inicial de documentos recuperados. Enquanto o recuperador inicial (bi-encoder) processa a consulta e os documentos separadamente para eficiência computacional, o re-rankeador avalia conjuntamente o par (consulta, documento), proporcionando pontuações de relevância muito mais precisas e contextualizadas, embora com maior custo computacional.

Compressão e Filtragem de Contexto: Técnicas avançadas que filtram, resumem ou comprimem inteligentemente as informações dos documentos recuperados para:

- ❑ Remover informações redundantes ou irrelevantes
- ❑ Extrair apenas os fatos mais pertinentes e informativos
- ❑ Reduzir o comprimento total do contexto para otimização computacional
- ❑ Manter a coerência e completude das informações essenciais

Self-Correction e Corrective RAG (CRAg): Abordagens recentes e sofisticadas que propõem pipelines cílicos e adaptativos onde um LLM especializado avalia criticamente a qualidade e relevância dos documentos recuperados. Se a informação for considerada irrelevante, incompleta ou baixa qualidade, o sistema pode automaticamente:

- ❑ Acionar uma nova busca com consulta refinada
- ❑ Buscar em fontes alternativas (como busca web em tempo real)
- ❑ Aplicar estratégias de recuperação diferentes
- ❑ Solicitar esclarecimentos adicionais do usuário

A segmentação textual (chunking) constitui uma etapa crítica e subestimada no pipeline RAG, definindo os blocos de conhecimento que serão recuperados e utilizados para geração. A estratégia de segmentação adotada impacta diretamente a qualidade, relevância e coerência das informações recuperadas, influenciando o desempenho final do sistema.

Tabela 1 – Comparativo entre Técnicas de Segmentação. (Elaborada pelo autor)

Técnica	Descrição	Vantagens	Desvantagens	Casos de Uso Ideais
Tamanho Fixo	Divide texto em segmentos de N caracteres com sobreposição. (LEWIS et al., 2020a)	<ul style="list-style-type: none"> • Implementação simples • eficiente • Controle preciso do tamanho 	<ul style="list-style-type: none"> • Ignora limites semânticos • Pode quebrar sentenças e ideias • Perda de coerência contextual 	<ul style="list-style-type: none"> • Prototipagem rápida • Recursos computacionais limitados • Textos homogêneos
Recursiva	Hierarquia de separadores (\n\n, \n, .., espaços) com fallback para tamanho fixo (ZHANG et al., 2024)	<ul style="list-style-type: none"> • Respeita estrutura textual • Flexibilidade adaptativa • Preserva parágrafos e sentenças 	<ul style="list-style-type: none"> • Chunks de tamanhos variados • Dependente da formatação • Pode gerar segmentos muito pequenos 	<ul style="list-style-type: none"> • Documentos bem formatados • Textos estruturados • Aplicações gerais
TextTiling	Detecta mudanças de subtópicos baseado em coesão lexical (HEARST, 1997)	<ul style="list-style-type: none"> • Fundamentação linguística • Segmentação temática • Eficaz para textos longos 	<ul style="list-style-type: none"> • Sensível a parâmetros • Ineficaz em textos curtos • Dependente de vocabulário 	<ul style="list-style-type: none"> • Documentos acadêmicos • Artigos longos • Textos expositivos
C99	Algoritmo baseado em matrizes de similaridade de coseno (CHOI; WIEMER-HASTINGS; MORE, 2000)	<ul style="list-style-type: none"> • Robusto e eficiente • Baseline acadêmico estabelecido • Boa performance geral 	<ul style="list-style-type: none"> • Foco apenas em coesão lexical • Perde nuances semânticas • Requer ajuste de parâmetros 	<ul style="list-style-type: none"> • Benchmarking • Comparações acadêmicas • Textos informativos
Baseada em Estrutura	Utiliza elementos estruturais explícitos (capítulos, seções, parágrafos) (RAGLC)	<ul style="list-style-type: none"> • Preserva integridade lógica • Mantém hierarquia original • Semanticamente coerente 	<ul style="list-style-type: none"> • Dependente de formatação • Limitada a textos estruturados • Pode gerar chunks muito grandes 	<ul style="list-style-type: none"> • Documentos técnicos • Manuais e guias • Textos acadêmicos
Semântica	Agrupa blocos baseado em similaridade de embeddings de sentenças (RAGLC)	<ul style="list-style-type: none"> • Coerência temática • Agrupa conteúdo relacionado • Independente de adjacência 	<ul style="list-style-type: none"> • custosa • Requer modelo de embedding • Complexidade de implementação 	<ul style="list-style-type: none"> • Bases de conhecimento complexas • Documentos heterogêneos • Aplicações de alta precisão

2.2.3.4 Frameworks de Avaliação Especializados

RAGAS (Retrieval Augmented Generation Assessment): Framework abrangente que fornece métricas especializadas para avaliação holística de sistemas RAG, incluindo avaliação automática e semi-automática de múltiplas dimensões de qualidade.

ARES (Automated RAG Evaluation System): Sistema automatizado que utiliza LLMs para avaliar a qualidade de sistemas RAG sem necessidade de ground truth manual, permitindo avaliação escalável e contínua.

2.2.3.5 Métricas Fundamentais

Context Precision: Mede a qualidade do recuperador avaliando se os chunks recuperados são genuinamente relevantes para responder à consulta:

$$\text{Context Precision}@k = \frac{\text{Chunks relevantes entre os top-k}}{k}$$

Context Recall: Avalia se todos os chunks relevantes disponíveis na base de conhecimento foram efetivamente recuperados:

$$\text{Context Recall} = \frac{\text{Chunks relevantes recuperados}}{\text{Total de chunks relevantes na base}}$$

2.2.3.6 Métricas Complementares

BLEU (PAPINENI et al., 2002) e ROUGE (LIN, 2004): Métricas tradicionais de Processamento de Linguagem Natural (PLN) que comparam a resposta gerada com respostas de referência (ground truth), medindo sobreposição de n-gramas e subsequências.

BERTScore (ZHANG et al., 2020): Métrica semântica que utiliza embeddings contextuais para avaliar similaridade semântica entre resposta gerada e referência, superando limitações lexicais das métricas tradicionais.

Similaridade de Cosseno de Embeddings: Medida direta de similaridade semântica no espaço vetorial entre resposta gerada e referência ideal.

2.3 Tecnologias e Ferramentas

A implementação bem-sucedida de sistemas RAG modernos e aplicações de NLP é viabilizada por um ecossistema tecnológico rico, diversificado e em constante evolução, composto por ferramentas especializadas, bibliotecas otimizadas, plataformas de desenvolvimento e infraestruturas computacionais avançadas. Esta seção apresenta uma análise abrangente e detalhada das principais tecnologias, frameworks e ferramentas que constituem o estado da arte atual no desenvolvimento de sistemas inteligentes baseados em LLMs e RAG.

2.3.1 Linguagens de Programação e Ambientes de Desenvolvimento

Python estabeleceu-se como a linguagem de programação predominante e praticamente universal no campo de IA e NLP, devido à sua sintaxe intuitiva, vasto ecossistema de bibliotecas especializadas, comunidade ativa e extensa documentação. A linguagem oferece bibliotecas fundamentais como NumPy para computação numérica eficiente, Pandas para manipulação de dados estruturados, e Matplotlib/Seaborn para visualização de dados e resultados experimentais.

Jupyter Notebooks e **Google Colab** fornecem ambientes interativos ideais para prototipagem rápida, experimentação iterativa e desenvolvimento colaborativo, permitindo combinação seamless de código, visualizações e documentação narrativa.

2.3.2 Frameworks de DL e Computação Neural

PyTorch emergiu como o framework de DL mais popular e amplamente adotado na comunidade de pesquisa, oferecendo uma Application Programming Interface (API) intuitiva e pythônica, execução dinâmica de grafos computacionais (eager execution), e ferramentas avançadas para debugging e visualização. Sua flexibilidade arquitetural e facilidade de customização o tornam ideal para pesquisa e desenvolvimento de modelos inovadores.

TensorFlow/Keras mantém relevância significativa, especialmente para aplicações de produção que requerem otimização de performance, deployment em escala e integração com infraestruturas empresariais. O TensorFlow Serving e TensorFlow Lite oferecem soluções especializadas para deployment em produção e dispositivos móveis, respectivamente.

JAX representa uma alternativa emergente que combina a familiaridade do NumPy com capacidades avançadas de diferenciação automática, compilação just-in-time (JIT) e paralelização automática, oferecendo performance superior para certas classes de problemas.

2.3.3 Plataformas de Modelos e Recursos Pré-Treinados

Hugging Face Hub constitui o repositório central e mais abrangente de modelos de linguagem pré-treinados, oferecendo acesso a milhares de modelos especializados, datasets curados e ferramentas de avaliação. A biblioteca Transformers fornece APIs unificadas e otimizadas para utilização de modelos de diferentes arquiteturas (BERT, GPT, T5, etc.) com código mínimo e máxima eficiência.

OpenAI API oferece acesso programático a modelos de linguagem de grande escala como GPT-4, GPT-3.5 e modelos especializados, através de interfaces REST bem

documentadas e SDKs oficiais para múltiplas linguagens de programação.

Anthropic Claude, **Google PaLM/Gemini** e **Cohere** fornecem alternativas competitivas com características distintivas, capacidades especializadas e modelos de pricing diferenciados.

2.3.4 Frameworks de Orquestração RAG

LangChain estabeleceu-se como o framework mais popular e abrangente para orquestração de pipelines RAG complexos, oferecendo abstrações de alto nível para integração de múltiplos componentes (LLMs, retrievers, memory systems, tools), templates de prompts sofisticados, e chains pré-construídas para casos de uso comuns. Sua arquitetura modular permite customização extensiva e integração com diversos provedores de LLM e bases de dados vetoriais.

LlamaIndex (anteriormente GPT Index) foca especificamente em indexação inteligente e recuperação eficiente de documentos, oferecendo estruturas de dados especializadas (list index, tree index, keyword table index) e algoritmos otimizados para diferentes tipos de consultas e documentos.

Haystack fornece uma alternativa robusta com foco em aplicações de produção, oferecendo pipelines modulares, componentes otimizados para performance e integrações nativas com ferramentas de monitoramento e observabilidade.

2.3.5 Bancos de Dados Vetoriais e Sistemas de Busca

ChromaDB oferece uma solução leve, embeddable e fácil de usar para armazenamento e busca de embeddings, ideal para prototipagem e aplicações de pequena a média escala. Sua API simples e integração nativa com Python o tornam popular para desenvolvimento inicial.

Pinecone fornece um serviço de banco de dados vetorial totalmente gerenciado, otimizado para performance e escala, com recursos avançados como filtragem de metadados, busca híbrida e monitoramento em tempo real.

Weaviate combina capacidades de banco de dados vetorial com funcionalidades de grafo de conhecimento, permitindo consultas complexas que combinam similaridade semântica com relações estruturadas.

Qdrant, **Milvus** e **FAISS** oferecem alternativas especializadas com diferentes trade-offs entre performance, escalabilidade e facilidade de uso.

2.3.6 Modelos de Embedding Especializados

Sentence Transformers fornece modelos pré-treinados otimizados especificamente para geração de embeddings de sentenças e documentos, com modelos especializados para diferentes idiomas e domínios.

OpenAI Embeddings (text-embedding-ada-002) oferece embeddings de alta qualidade através de API, com performance consistente em múltiplas tarefas e idiomas.

Modelos Multilíngues como `intfloat/multilingual-e5-large` demonstram performance superior em aplicações que requerem suporte a múltiplos idiomas simultaneamente, sendo relevantes para contextos brasileiros que podem envolver português, inglês e outras linguagens.

2.3.7 Ferramenta de Interface

FastAPI fornece um framework moderno e de alta performance para criação de APIs REST, com documentação automática, validação de tipos e suporte nativo a operações assíncronas.

2.3.8 Plataformas de Automação e Integração: N8N

O **N8N** representa uma das plataformas de automação de fluxo de trabalho mais inovadoras e poderosas disponíveis no ecossistema tecnológico contemporâneo, constituindo-se como uma ferramenta fundamental para orquestração complexa de processos, integração de sistemas heterogêneos e automação inteligente de tarefas em implementações de sistemas RAG e aplicações de IA. Esta plataforma **opensource** para utilização não comercial, desenvolvida com foco em extensibilidade, usabilidade e flexibilidade arquitetural, oferece capacidades extraordinárias para conectar, sincronizar e automatizar interações entre múltiplos serviços, APIs e sistemas de dados, incluindo a integração com aplicativos de mensagens como WhatsApp e Telegram.

2.3.9 APIs de Comunicação Avançada: Evolution API

Utilizada nesse trabalho apenas para fins acadêmicos, A **Evolution API** representa uma solução tecnológica altamente especializada para integração avançada com o ecossistema WhatsApp Business, constituindo-se como uma ferramenta fundamental para implementação de sistemas conversacionais inteligentes que operam através da plataforma de mensageria mais popular globalmente. Esta API, totalmente **opensource** para fins não comerciais e criada por um Brasileiro, oferece capacidades extraordinárias para automação de comunicações, integração de chatbots inteligentes e desenvolvimento de soluções de atendimento automatizado que mantêm a naturalidade e eficácia da comunicação humana.

2.4 Configuração Utilizada (*setup*)

Para as implementações práticas e aplicação das pesquisas no âmbito deste trabalho, foram utilizada as configurações tecnológicas (recomendadas) com licenças **opensource**

para utilização própria ou de baixo custo, que incluem:

- ❑ **Modelo de Linguagem Central:** DeepSeek-V3 selecionado com base nos benchmarks publicados e no custo para utilização, sendo um dos modelos com melhor custo/benefício da atualidade
- ❑ **Modelo de Embedding:** intfloat/multilingual-e5-large para aplicações em leis, editais e instruções normativas de língua portuguesa, conforme testes especificados no próximo capítulo.
- ❑ **Banco de Dados Vetorial:** ChromaDB devido a facilidade de instalação e utilização, seu custo e sua performance comprovadas por meio de benchmarks.
- ❑ **Framework de Orquestração e automação de fluxos:** N8N devido ao fato de ser uma ferramenta opensource e lowcode, permitindo que os próprios usuários possam facilmente adaptar o projeto às necessidades e particularidades de cada processo.
- ❑ **Interface de Comunicação:** Evolution API para integração WhatsApp Business em razão de ser uma plataforma *opensource*, no caso do atual utilizada para fins acadêmicos, pela sua aceitação e facilidade de uso dentro da comunidade, e principalmente por sua aderência à orquestração no N8N.
- ❑ **Interface de Recuperação por Contexto:** Foi utilizado o FastAPI para criação de uma API Python onde é possível enviar um arquivo no formato PDF e ter o retorno da quebra desse arquivo em unidades lógicas contextuais retornadas em forma de um arquivo JSON. Este arquivo é utilizado para a criação dos embeddings e posterior formação do contexto para o LLM no RAG.
- ❑ **COLAB:** Essa ferramenta disponibilizada pelo google foi usada em vários experimentos e comprovações ao longo deste projeto.

Esta configuração tecnológica integrada e abrangente permitiu o desenvolvimento eficiente, a experimentação rigorosa e o deployment confiável de um sistema RAG avançado dentro do aplicativo de mensagens WhatsApp, o qual é muito eficaz na comunicação com os servidores (como mostra um dos experimentos realizados) e aplicável especialmente em ambientes de administração pública onde acessibilidade, confiabilidade e escalabilidade são requisitos fundamentais.

2.5 Trabalhos Relacionados e Contexto Científico

A aplicação de Grandes Modelos de Linguagem e técnicas de RAG no setor público e em domínios especializados representa um campo de pesquisa emergente e promissor, com

contribuições científicas crescentes que abordam tanto aspectos técnicos quanto desafios práticos. Esta seção apresenta uma revisão crítica e abrangente dos trabalhos mais relevantes e influentes na área, contextualizando as contribuições científicas contemporâneas e identificando lacunas de pesquisa que motivam investigações futuras.

2.5.1 Aplicações de LLMs no Setor Público

Kleiman e Barbosa (2024) descrevem o desenvolvimento e implementação de um chatbot denominado ChatPGD (KLEIMAN; BARBOSA, 2024), um sistema de chatbot inteligente especificamente projetado para responder a dúvidas e consultas sobre o Programa de Gestão e Desempenho do governo federal brasileiro. Este trabalho pioneiro utiliza o modelo GPT-3.5 como base tecnológica e enfrenta desafios similares aos abordados em pesquisas contemporâneas, incluindo integração eficaz de dados governamentais heterogêneos, garantia de precisão e confiabilidade das respostas, e manutenção de conformidade com regulamentações específicas do setor público. O objetivo principal foi investigar os benefícios e os riscos do uso de um chatbot baseado em ChatGPT para automatizar as respostas relacionadas ao PGD.

A Pesquisa com Usuários foi aplicada através de um questionário a uma comunidade de mais de 2.100 servidores públicos que testaram a ferramenta, obtendo 85 respostas. A pesquisa avaliou a utilidade, a precisão das informações e o desempenho técnico do sistema. Em um período de 15 dias foram 1985 mensagens processadas em 724 conversas distintas.

Apesar do trabalho não detalhar a arquitetura utilizada, menciona que não se tratou de um "fine-tuning" tradicional e teve acesso a uma base de dados para consulta em tempo real

Os pesquisadores obtiveram resultados importantes sendo que, a maioria dos usuários considerou a ferramenta útil. 72% dos respondentes da pesquisa afirmaram que o ChatPGD foi útil para esclarecer suas dúvidas sobre o programa. Foram identificadas falhas de precisão, onde 18% dos usuários relataram ter recebido informações incorretas ou irrelevantes e em relação ao desempenho, o resultado foi geralmente considerado positivo, com 73% dos usuários não relatando problemas técnicos. No entanto, 18% enfrentaram mensagens de erro e 9% notaram lentidão no sistema.

O estudo de Kleiman e Barbosa foca primordialmente em uma abordagem de recuperação de informações relativamente simples e direta, não explorando sistematicamente otimizações avançadas de segmentação de documentos, técnicas sofisticadas de embeddings contextuais, ou estratégias de enriquecimento de consultas que têm demonstrado melhorias significativas em sistemas RAG contemporâneos.

Esta lacuna representa uma oportunidade importante para pesquisas futuras que busquem otimizar tecnicamente o desempenho de sistemas conversacionais governamentais.

2.5.2 Governo Algorítmico e IA Pública

O trabalho abrangente e visionário de Engin e Treleaven (ENGIN; TRELEAVEN, 2018) (2018) fornece um panorama teórico amplo e multidisciplinar sobre o conceito emergente de "governo algorítmico", discutindo sistematicamente a integração sinérgica de tecnologias de IA, Internet das Coisas (IoT) e blockchain para automatizar, otimizar e transformar a prestação de serviços públicos. Esta pesquisa oferece uma perspectiva conceitual e estratégica valiosa, explorando implicações sociopolíticas, desafios éticos e oportunidades transformacionais da digitalização governamental.

Entretanto, a abordagem de Engin e Treleaven é predominantemente conceitual e de revisão bibliográfica, não apresentando implementações empíricas concretas, validações experimentais ou estudos de caso detalhados que demonstrem a viabilidade prática e eficácia real das tecnologias propostas em contextos governamentais específicos. Esta limitação destaca a necessidade de pesquisas que combinem rigor teórico com validação empírica robusta.

2.5.3 Retrieval-Augmented Generation for Large Language Models: A Survey

O trabalho de Gao et al. (GAO et al., 2024) apresenta uma revisão abrangente sobre a Geração Aumentada por Recuperação (RAG), uma técnica que aprimora os Grandes Modelos de Linguagem (LLMs) ao integrar conhecimento de fontes de dados externas. Esta abordagem surge como uma solução promissora para desafios inerentes aos LLMs, como a geração de informações incorretas (alucinações), conhecimento desatualizado e a falta de transparência nos processos de raciocínio. Ao combinar o conhecimento paramétrico dos LLMs com repositórios de dados externos e dinâmicos, o RAG aumenta a precisão e a credibilidade das respostas, especialmente em tarefas que exigem conhecimento intensivo, permitindo atualizações contínuas e a integração de informações de domínios específicos.

A pesquisa detalha a evolução do RAG através de três paradigmas principais: Naive RAG, RAG Avançado (Advanced RAG) e RAG Modular (Modular RAG). Inicialmente, o Naive RAG, que ganhou destaque com a adoção generalizada do ChatGPT, segue um processo tradicional de indexação, recuperação e geração. No entanto, este paradigma inicial enfrenta limitações significativas, como desafios na precisão e recall da recuperação, dificuldades na geração (alucinações, irrelevância) e obstáculos na integração de informações, que podem levar a saídas incoerentes ou redundantes.

Para superar as deficiências do Naive RAG, o RAG Avançado introduz melhorias específicas, focando na otimização da qualidade da recuperação. Isso é alcançado por meio de estratégias pré-recuperação e pós-recuperação. As técnicas pré-recuperação incluem aprimoramento da granularidade dos dados, otimização das estruturas de índice, adição de metadados e otimização de consultas. Já as técnicas pós-recuperação envolvem a

reclassificação de chunks e a compressão de contexto, garantindo que as informações mais relevantes sejam priorizadas e que a sobrecarga de informações seja mitigada antes da geração da resposta pelo LLM.

O RAG Modular representa a arquitetura mais avançada, oferecendo adaptabilidade e versatilidade aprimoradas. Este paradigma permite a introdução de módulos funcionais específicos e a substituição de módulos existentes, resultando em maior flexibilidade. Diferente dos paradigmas anteriores, o RAG Modular não se limita a processos sequenciais de recuperação e geração, incorporando métodos como recuperação iterativa e adaptativa. Novos módulos, como o módulo de Busca (para pesquisa direta em diversas fontes de dados), o módulo de Memória (para guiar a recuperação com base na memória do LLM) e o módulo de Adaptação de Tarefas (para personalizar o RAG para tarefas específicas), são introduzidos para otimizar as capacidades de recuperação e processamento.

O artigo também compara o RAG com outras metodologias de otimização de modelos, como o Fine-tuning (FT) e a Engenharia de Prompt. Enquanto a Engenharia de Prompt exige poucas modificações no modelo e conhecimento externo, focando em aproveitar as capacidades intrínsecas dos LLMs, o Fine-tuning envolve um treinamento adicional do modelo. O RAG se destaca por sua capacidade de fornecer atualizações de conhecimento em tempo real e utilizar efetivamente fontes de conhecimento externas, com alta interpretabilidade, sendo ideal para ambientes dinâmicos. Embora o FT permita uma personalização profunda do comportamento e estilo do modelo, ele é mais estático e exige retreinamento para atualizações, além de ser computacionalmente intensivo. O estudo ressalta que RAG e FT não são mutuamente exclusivos e podem ser combinados para otimizar o desempenho.

O trabalho também analisameticulosamente os componentes fundamentais do RAG: Recuperação (Retrieval), Geração (Generation) e Aumento (Augmentation). Na fase de Recuperação, são abordadas questões como a fonte dos dados (estruturados, semi-estruturados, não-estruturados e conteúdo gerado por LLMs), a granularidade da recuperação (frase, sentença, chunk, documento) e a otimização do processo de indexação e consulta. A fase de Geração foca em como o LLM formula a resposta a partir do contexto recuperado, incluindo técnicas de fine-tuning e pós-processamento. Por fim, a fase de Aumento explora como a informação recuperada é integrada e utilizada para aprimorar a saída do LLM, abordando desafios como a coerência e a relevância.

Além disso, o artigo apresenta um framework de avaliação atualizado e benchmarks para o RAG, cobrindo 26 tarefas e quase 50 conjuntos de dados. Ele descreve os objetivos e métricas de avaliação, bem como as ferramentas e benchmarks existentes. Esta seção é crucial para pesquisadores e desenvolvedores que buscam avaliar e comparar o desempenho de diferentes sistemas RAG de forma sistemática e rigorosa, garantindo a reproduzibilidade e a validade dos resultados em diversas aplicações e cenários.

Finalmente, o estudo delineia os desafios atuais enfrentados pelo RAG e aponta ave-

nidas promissoras para pesquisa e desenvolvimento futuros. Estes incluem a necessidade de lidar com a complexidade crescente dos dados, aprimorar a robustez e a eficiência dos sistemas RAG em cenários do mundo real, e explorar novas arquiteturas e técnicas de otimização. O artigo visa fornecer uma compreensão detalhada e estruturada dos conceitos técnicos fundamentais, da progressão histórica e do espectro de metodologias e aplicações do RAG, servindo como um guia essencial para pesquisadores e profissionais da área.

2.5.4 Posicionamento Científico e Contribuições Distintivas

As pesquisas contemporâneas se posicionam estratégicamente tentando preencher lacunas importantes ao focar na otimização técnica sistemática e rigorosa do pipeline RAG (incluindo segmentação avançada, embeddings especializados, e enriquecimento intelectual de consultas) dentro de casos de uso práticos, porém, poucos trabalhos tem como foco a Administração Pública e a maioria, como é notório no estado da arte proposto por .

2.6 Considerações Finais da Fundamentação Teórica

A fundamentação teórica apresentada neste capítulo estabelece um arcabouço conceitual abrangente e rigoroso que sustenta o desenvolvimento de sistemas inteligentes baseados em Grandes Modelos de Linguagem e técnicas de RAG. A evolução paradigmática desde abordagens clássicas de NLP até as arquiteturas sofisticadas contemporâneas demonstra a maturidade crescente do campo e o potencial transformador dessas tecnologias.

A análise detalhada das arquiteturas Transformer, dos mecanismos de atenção e das técnicas de representação vetorial revela a complexidade técnica subjacente aos avanços recentes em IA, enquanto a discussão sobre RAG e suas otimizações ilustra como a integração sinérgica de diferentes paradigmas pode superar limitações individuais de cada abordagem. As técnicas de segmentação de texto, metodologias de avaliação e frameworks tecnológicos apresentados fornecem as ferramentas práticas necessárias para implementações eficazes e robustas.

O panorama de trabalhos relacionados e o ecossistema tecnológico atual demonstram tanto as oportunidades quanto os desafios presentes no campo, indicando direções promissoras para pesquisas futuras e aplicações práticas. A convergência entre rigor teórico, inovação técnica e aplicabilidade prática posiciona as tecnologias discutidas como elementos centrais na transformação digital de diversos setores, incluindo a administração pública e serviços governamentais.

Esta fundamentação teórica sólida e abrangente estabelece as bases conceituais necessárias para o desenvolvimento, implementação e avaliação de sistemas inteligentes que

possam contribuir para a melhoria da eficiência, acessibilidade e qualidade dos serviços públicos, representando um passo importante na direção de uma administração pública mais inteligente, responsiva e centrada no cidadão.

CAPÍTULO 3

Metodologia Proposta e Desenvolvimento

Este capítulo detalha a metodologia no desenvolvimento e avaliação da solução para automatizar o atendimento a candidatos de concursos e processos seletivos e ingressantes aos cargos da Prefeitura Municipal de Uberlândia. A proposta, denominada **RAGLC** (Retrieval Augmented Generation in Logical Context), visa formalizar um modelo robusto baseado em NLP e Modelos de Linguagem de Grande Escala (LLMs), com foco em técnicas de RAG, porém, buscando técnicas inovadoras que demonstraram excelentes resultados na melhoria da formação de contexto, melhorando a resposta emitida pelo *LLM*, tudo funcionando de forma acessível dentro do aplicativo de mensagens WhatsApp.

O objetivo de criar um modelo capaz de interpretar perguntas em linguagem natural e fornecer respostas precisas, utilizando como base de conhecimento os editais e documentos correlatos. As contribuições efetivas residem na exploração de métodos de segmentação contextual e na otimização da recuperação de informações através de abordagens híbridas.

3.1 Visão Geral do Pipeline Experimental

A pesquisa foi estruturada em um pipeline sequencial e iterativo, conforme ilustrado abaixo.

1. Definição do Problema e Coleta de Dados: Identificação das maiores demandas do Núcleo de Recrutamento e coleta dos documentos base (editais de concursos e leis municipais alusivas aos servidores públicos municipais da Prefeitura Municipal de Uberlândia).
2. Validação Preliminar relativa ao canal de comunicação (WhatsApp).
3. Análise qualitativa das respostas: Testes do modelo treinado com fine-tuning e do modelo RAG simples, ambos baseados na Lei Complementar nº 40/1992.

4. Geração de Embeddings: Teste de múltiplos modelos geradores de embeddings e bancos de dados vetoriais.
5. Pré-processamento e Segmentação: Construção do algoritmo, aplicação e comparação de diferentes estratégias de chunking (tradicional vs. contexto) com e sem enriquecimento contextual da pergunta do usuário utilizando métricas de qualidade (BERTSCORE, BLEU, etc.).
6. Síntese e Proposta de Trabalho: Análise dos resultados, conclusão e delineamento da arquitetura para o RAG Híbrido (RAGLC).

3.2 Definição do Problema e Coleta de Dados

Nessa fase inicial do projeto, a equipe do Núcleo de Recrutamento de Pessoal, da Prefeitura Municipal de Uberlândia, ligada à Secretaria Municipal de Administração foi indagada sobre suas demandas sazonais e recorrentes, ocasionadas por uma demanda altíssima por profissionais, trabalhando com prazos mínimos para atendimento a essas demandas e uma taxa de desistência de candidatos superior a 35Após o levantamento junto a repartição foram elencados alguns gargalos, principalmente em relação a respostas aos candidatos e ingressantes no serviço público municipal em razão do volume de dúvidas e orientações demandadas pelos candidatos via aplicativo de mensagens WhatsApp, chegando em alguns momentos a mais de 200 interações sem resposta em razão de ter que priorizar outras atividades mais urgentes. A ideia central proposta por esse trabalho visa aliviar a carga de atividades e trabalho do setor e prestar um atendimento mais satisfatório aos candidatos e recém admitidos da municipalidade.

3.3 Validação Preliminar: Canal de Comunicação

Antes de focar na AI da ferramenta, foi realizada uma avaliação preliminar para comparar a eficácia do uso de aplicativos de mensagens instantâneas (WhatsApp) em relação ao e-mail como canal de comunicação com os candidatos. Esta etapa visou entender a preferência e a taxa de resposta dos usuários, justificando a escolha do canal para futuras interações do RAGLC. Já havia sido feito um estudo técnico preliminar pela própria Prefeitura, a fim de verificar a eficácia de ligações telefônicas na comunicação com servidores e ingressantes e de fato, pelo resultado dessa pesquisa, foi possível verificar a sua ineeficácia no contexto atual.

Foram selecionados 22 candidatos aprovados no concurso público nº 01/2023 da Prefeitura Municipal de Uberlândia. Mensagens de convocação padronizadas foram enviadas para este grupo tanto por e-mail, quanto por aplicativo de mensagens WhatsApp. As seguintes métricas foram coletadas e comparadas:

- ❑ **Tempo médio de resposta dos usuários:** Calculou-se o intervalo entre o envio da mensagem e a primeira resposta ou confirmação do candidato.
- ❑ **Taxa de resposta:** Percentual de candidatos que responderam dentro do prazo de 48 horas.
- ❑ **Retorno qualitativo da satisfação (informal):** Coleta de feedback espontâneo dos candidatos sobre a conveniência do canal utilizado.

Os resultados desta avaliação preliminar (detalhados no Capítulo 5) indicaram uma preferência e maior agilidade de resposta via WhatsApp, suportando a decisão de explorar este canal para a interface do presente projeto, embora a tecnologia principal (RAGLC) seja independente do canal de entrega, utilizar a estrutura do aplicativo de mensagens torna a ferramenta mais acessível, tendo em vista a grande utilização e aceitação do aplicativo tanto no setor de Recrutamento, quanto junto aos candidatos.

3.4 Análise da melhor estratégia para resolução do problema específico

Nesta fase, foram testadas duas estratégias para a obtenção das respostas junto ao modelo LLM. A primeira estratégia envolveu o treinamento de modelos baseados no Llama. Para esse treinamento foram preparados alguns datasets com perguntas e respostas variando entre 200 a 500 perguntas extraídas da Lei Complementar nº 40/1992 (Estatuto dos servidores públicos municipais). Desse processo, nasceu o modelo LEI40_model, desenvolvido através de um processo de fine-tuning de adaptação de domínio (domain adaptation fine-tuning) a partir do modelo base Meta-Llama-3.1-8B e treinado com uma base de dados de 500 perguntas e respostas sobre a Lei e com outra base de dados com a própria Lei segmentada em seus artigos. Para otimizar o processo e torná-lo computacionalmente eficiente, foi empregada uma técnica de Parameter-Efficient Fine-Tuning (PEFT), especificamente o método Q-LoRA (Quantized Low-Rank Adaptation), que realiza o ajuste fino utilizando uma versão do modelo com parâmetros quantizados em 4-bits. Esse modelo foi posteriormente disponibilizado a toda a comunidade via huggingface. A segunda estratégia foi criar um modelo do tipo Naive RAG (LEWIS et al., 2021), utilizando um modelo mais simples da OpenAI via API e um modelo de embedding também compatível com a mesma. Após a produção das duas estratégias, ambas foram submetidas a 20 perguntas formuladas por dois integrantes do Núcleo Jurídico da Secretaria Municipal de Administração, os quais também participaram da análise das respostas dos modelos. O RAG, mesmo em uma estrutura muito simples obteve um resultado absolutamente superior ao modelo treinado. Esse fenômeno se deu em razão da dificuldade de obtenção de dados para treinamento do modelo, tendo em vista que a base utilizada para

isso foi apenas a Lei Complementar nº 40/1992. Este experimento, demonstrou que, para o escopo atual do trabalho, tendo em vista que são utilizados datasets pontuais específicos em curtos espaços de tempo, não era possível utilizar a primeira estratégia, enquanto a segunda, embora se tratando de uma implementação simples, já obteve resultados muito satisfatórios.

3.5 Avaliação de Modelos de Embeddings

A qualidade dos embeddings é fundamental para o desempenho de qualquer sistema RAG, pois influencia diretamente a capacidade de encontrar os chunks de texto mais relevantes para uma dada consulta. Reconhecendo isso, foi realizada uma investigação comparativa entre diferentes modelos de embeddings, focando em sua adequação ao domínio específico de editais e textos administrativos/jurídicos em português.

A proposta foi a criação de um algoritmo baseado na estrutura de algoritmos genéticos (HOLLAND, 1973), testando vários hiperparâmetros do AG como forma de avaliar diferentes modelos disponíveis no âmbito específico abordado por esse trabalho, sendo eles os seguintes: *neuralmind/bert-base-portuguese-cased* (SOUZA; NOGUEIRA; LOTUFO, 2020), *neuralmind/bert-large-portuguese-cased* (SOUZA; NOGUEIRA; LOTUFO, 2020), *intfloat/multilingual-e5-large* (WANG et al., 2024), *sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2* (REIMERS; GUREVYCH, 2019c) e *BAAI/bge-large-en-v1.5* (XIAO et al., 2023).

A metodologia de avaliação consistiu em:

1. Utilizar o mesmo conjunto de chunks (segmentados contextual ou semanticamente) e o mesmo conjunto de perguntas de teste (com respostas mapeadas para os chunks corretos - *ground truth*).
2. Gerar os embeddings para os chunks e as perguntas usando cada um dos cinco modelos.
3. Para cada pergunta, realizar uma busca por similaridade (usando similaridade de cosseno) no banco de embeddings dos chunks gerados por um modelo específico.
4. Foi medida a acurácia de cada modelo, verificando se o(s) chunk(s) correto(s) estava(m) entre os k resultados mais similares recuperados (variando $k = 1, 2, 3$).

A análise comparativa focou não apenas na acurácia bruta, mas também em observações qualitativas sobre quais modelos pareciam capturar melhor as nuances semânticas do domínio.

Os resultados (detalhados no Capítulo 5) confirmaram que a escolha do modelo de embedding tem um impacto no desempenho da recuperação. O modelo *intfloat/multilingual-e5-large*, pré-treinado em grandes volumes de dados e que opera em níveis contextuais (não

apenas palavras isoladas), demonstrou um desempenho surpreendente, sendo superior inclusive a modelos de grande escala treinados exaustivamente em língua portuguesa. A capacidade de gerar representações vetoriais que capturam o significado contextual de cláusulas e seções inteiras de editais e leis foi crucial. Esta análise fornece uma contribuição prática ao identificar modelos de embedding eficazes para aplicações de RAG em contextos legais em língua portuguesa, orientando futuros trabalhos na área.

Foram testados, aproveitando o mesmo algoritmo, 3 bancos de dados vetoriais (FAISS, ChromaDB e QDrant), onde foi avaliada a performance em relação ao contexto, a utilização de memória e a velocidade de retorno dos embeddings.

3.6 Algoritmo de Segmentação por Similaridade Semântica

Explorando alternativas à segmentação tradicional por número de caracteres e à segmentação puramente estrutural (capítulos/seções/itens/subitens), foi desenvolvido e avaliado um algoritmo de segmentação baseado em blocos semânticos, inspirado na estrutura de textos legais e editais de concursos ou processos seletivos do município.

A motivação partiu da hipótese de que, para documentos com forte estrutura lógica e semântica como leis ou editais, com dificuldade de acurácia conhecida (HUANG et al., 2024), agrupar o conteúdo por unidades temáticas coesas poderia melhorar a qualidade do contexto recuperado pelo RAG.

O algoritmo proposto funciona da seguinte forma:

- Segmentação Estrutural Inicial:** O documento (Leis Municipais e Editais) é dividido com base em delimitadores naturais ("Art.", "Seção", "CAPÍTULO", "Itens", Parágrafos). Cada bloco inicial representa uma unidade estrutural.
- Geração de Embeddings:** Embeddings são gerados para cada bloco inicial usando um modelo de melhor performance nessa tarefa específica.
- Agrupamento Semântico:** A similaridade (cosseno) entre os embeddings de todos os blocos é calculada. Blocos cuja similaridade ultrapassa um limiar pré-definido são agrupados, formando "blocos semânticos" maiores que reúnem conteúdo tematicamente relacionado, mesmo que não estejam adjacentes no texto original.
- Indexação e Recuperação RAG:** Estes blocos semânticos agrupados (ou os blocos estruturais originais, se não agrupados) são indexados em um banco de dados vetorial. No momento da consulta, a pergunta do usuário é convertida em embedding, e os k blocos semânticos mais similares são recuperados para compor o contexto do LLM *DeepSeek* (LIU et al., 2024)), utilizado nesse experimento.

Testes massivos foram produzidos a fim de que fosse avaliado se esta abordagem de "*RAG Semântico*" produziria respostas de melhor qualidade em comparação com:

- **RAG Tradicional:** Segmentação por tamanho fixo de caracteres com overlap.
- **RAG segmentado por Bloco, sem agrupamento por semelhança semântica:** Segmentação baseada nos delimitadores naturais, sem agrupamento semântico.

Primeiramente os experimentos comparativos foram realizados com um conjunto de perguntas sobre o Estatuto dos Servidores Municipais e sobre o Edital de Processo Seletivo nº 01/2025, onde o primeiro teve suas respostas avaliadas por meio de Context Recall (ES et al., 2025).

Cada um dos três algoritmos RAG recebeu as mesmas 20 perguntas e utilizou o LLM DeepSeek V.3 para gerar as respostas. Essas respostas foram apreciadas pela Assessoria Jurídica da Secretaria Municipal de Administração, órgão responsável pela política de Administração de Pessoal no município de Uberlândia (órgão Gestor do Estatuto dos Servidores). Esta equipe avaliou a qualidade das respostas em uma escala de 0 a 10 onde 0 representaria uma resposta completamente incorreta e 10 uma resposta plenamente correta e fundamentada.

A partir do estudo propiciado por essa análise o algoritmo RAGLC alcançou uma acurácia de 98%, cinco pontos percentuais a mais que o modelo tradicional Naive RAG.

Um segundo experimento foi realizado, porém, utilizando um edital específico, o Edital nº 01/2025. A RBO Assessoria Pública e Projetos Municipais, organizadora de processos seletivos para a Prefeitura de Uberlândia, disponibilizou uma base de dados anonimizada contendo perguntas reais feitas por candidatos do Processo Seletivo retomencionado e as respectivas respostas fornecidas pela banca.

Com essa valiosa informação, foram introduzidas na ferramenta as mesmas perguntas reais dos candidatos exatamente como foram feitas junto a banca. Dessa vez o cenário testado envolveu também a manipulação (enriquecimento semântico ambientado) da pergunta feita pelo usuário visando entender se o resultado seria melhor nessas condições.

O procedimento foi o seguinte:

1. As perguntas reais da base de dados foram submetidas a ferramenta (que testou 6 abordagens diferentes de obtenção de contexto: Naive RAG, RAG por blocos e RAG por bloco e contexto, sendo testadas com e sem o enriquecimento semântico da pergunta do usuário).
2. As respostas geradas por todos os modelos foram coletadas.
3. As respostas coletadas foram comparadas lado a lado com as respostas oficiais da RBO.

As métricas de avaliação (BERTSCORE, COSINE_ST, F1_OVERLAP, JACCARD, BLEU, LEVENSHTEIN, COSINE_TFIDF, SCORE_PONDERADO_GERAL) foram selecionadas para fornecer uma visão multidimensional do desempenho do sistema, cobrindo aspectos de similaridade lexical, semântica e estrutural entre as respostas geradas e as respostas esperadas. A justificativa estatística para a inclusão de cada métrica é a seguinte:

Métricas Baseadas em Embeddings (BERTSCORE, COSINE_ST): São cruciais para capturar a similaridade semântica, que é frequentemente mais importante do que a similaridade lexical em tarefas de Geração de Linguagem Natural. Elas são robustas a variações de vocabulário e paráfrases, refletindo melhor a compreensão do significado. Estatisticamente, elas fornecem uma medida contínua de similaridade em um espaço vetorial de alta dimensão.

Métricas Baseadas em Overlap Lexical (F1_OVERLAP, JACCARD, BLEU): Complementam as métricas semânticas ao avaliar a presença de termos-chave e a correspondência exata de palavras. Embora possam ser menos sensíveis a paráfrases, são importantes para garantir que o sistema esteja utilizando o vocabulário correto e relevante do domínio. O BLEU, em particular, é útil para avaliar a fluidez e a gramaticalidade, embora sua aplicação principal seja em tradução automática.

Métrica de Distância de Edição (LEVENSHTEIN): Fornece uma medida de quão diferentes duas strings são em termos de operações de edição. É útil para identificar pequenas variações textuais e erros de digitação ou formatação.

Métrica Baseada em Frequência de Termos (COSINE_TFIDF): Oferece uma perspectiva sobre a similaridade baseada na importância estatística das palavras. É uma baseline clássica que ajuda a contextualizar o desempenho de modelos mais avançados.

SCORE_PONDERADO_GERAL: A inclusão de um score ponderado geral é estatisticamente justificada como uma forma de agregar múltiplas dimensões de qualidade em um único indicador. A ponderação de cada métrica neste score deve ser determinada com base na sua importância relativa para o objetivo final do sistema (maior peso para factualidade e relevância semântica). Este score facilita a comparação global entre as diferentes abordagens e a identificação da configuração de melhor desempenho.

A validação dessas métricas, especialmente o BERTSCORE, foi realizada para garantir que seu comportamento seja consistente e que eventuais "scores idênticos" (como mencionado nas diretrizes) sejam investigados e corrigidos, possivelmente através de uma análise mais granular ou da inclusão de outras métricas que capturem nuances não detectadas.

A escolha de um conjunto diversificado de métricas permite uma avaliação robusta e evita a dependência excessiva de um único indicador, que pode não capturar a complexidade total da qualidade da resposta.

Para o modelo baseado na obtenção de chunks diversificados com agrupamento semântico foram utilizados os hiperparâmetros abaixo no experimento 2 da subseção anterior.

min_similarity (Límiar de Similaridade): Este hiperparâmetro define a força da conexão semântica necessária para que dois blocos de texto sejam considerados parte da mesma comunidade no grafo de similaridade. A escolha de um valor (0.55) é baseada em experimentação preliminar e na literatura que sugere que valores entre 0.5 e 0.7 geralmente funcionam bem para similaridade de cosseno em embeddings de sentenças. Um valor muito baixo pode agrupar blocos semanticamente distantes, introduzindo ruído. Um valor muito alto pode fragmentar blocos semanticamente relacionados, resultando em chunks muito pequenos.

min_cluster_size (Tamanho Mínimo do Cluster): Este parâmetro garante que os chunks gerados tenham um tamanho mínimo de blocos. A escolha de `min_cluster_size = 2` visa evitar a criação de chunks com apenas um bloco, a menos que este bloco seja semanticamente isolado. Isso ajuda a manter uma certa densidade de informação em cada chunk, o que é benéfico para a fase de geração do RAG. A justificativa estatística aqui reside em evitar a criação de unidades de informação excessivamente granulares que poderiam prejudicar a capacidade do LLM de sintetizar respostas coerentes.

max_tokens_per_chunk (Máximo de Tokens por Chunk): Embora o DocumentChunker priorize a coerência semântica, um limite máximo de tokens é essencial para gerenciar a entrada do LLM. LLMs têm janelas de contexto limitadas, e chunks muito grandes podem exceder essa capacidade, levando à truncagem ou à perda de informações. A definição de um `max_tokens_per_chunk` (que será determinado com base na janela de contexto do LLM utilizado, DeepSeek) garante que os chunks sejam processáveis pelo modelo, enquanto a lógica de quebra sequencial dentro do DocumentChunker tenta preservar a ordem original e minimizar a fragmentação semântica quando o limite é atingido. A escolha deste parâmetro é uma trade-off entre a preservação da coerência semântica e as restrições computacionais e de modelo.

3.7 Dimensões de Avaliação da Proposta

Além das métricas técnicas de acurácia e desempenho, a avaliação da proposta pelo modelo RAGLC considerou dimensões mais amplas relacionadas ao seu impacto potencial no serviço público:

- ❑ **Benefícios da Automação:** A tecnologia proposta é realmente benéfica neste contexto específico de atendimento a candidatos? Reduz o tempo de espera? Aumenta a disponibilidade do serviço (24/7)?
- ❑ **Contexto de Aplicação:** Quais tipos de perguntas são mais adequadas para o chatbot? Onde a intervenção humana ainda é necessária?

- ❑ Possíveis Prejuízos:** Existem riscos associados? Respostas incorretas (alucinações)? Exclusão digital de usuários menos familiarizados com tecnologia? Custos de implantação e manutenção?
- ❑ Melhoria do Processo e Atendimento:** A automação efetivamente melhora o processo interno do Núcleo de Recrutamento? Reduz a carga de trabalho dos servidores? Aumenta a satisfação geral dos candidatos (usuários)? Contribui para a transparência?

Estas questões foram abordadas através da combinação dos resultados dos experimentos quantitativos (acurácia, tempo, usabilidade) e das análises qualitativas (comparação com respostas oficiais, feedback do corpo jurídico), buscando fornecer uma visão holística do valor e das limitações da solução proposta.

Os resultados indicaram que a abordagem de agrupamento semântico (RAGLC) obteve as melhores pontuações, para perguntas complexas que exigiam a correlação de informações de diferentes partes do documento. A capacidade de agrupar conteúdo similar seja qual for a sua posição no texto, fornece uma janela de contexto mais rico e completo para o LLM. Os resultados quantitativos e qualitativos desta comparação são detalhados no Capítulo 5, corroborando a validade desta abordagem.

CAPÍTULO 4

RAG Híbrido para Integração de Dados no Atendimento Público

4.1 Introdução

A informação no setor público raramente reside em um único formato. Enquanto editais, leis e portarias são predominantemente textuais (documentos não estruturados), dados cruciais sobre cidadãos, processos, e recursos frequentemente se encontram em bancos de dados relacionais, planilhas ou sistemas legados (dados estruturados). Um cidadão buscando informações sobre um concurso pode precisar saber não apenas as regras do edital (texto), mas também o status atual de sua inscrição, a classificação parcial, ou o número de vagas já preenchidas (dados estruturados). Sistemas RAG convencionais, focados em recuperação de texto, são incapazes de responder a essas perguntas complexas que exigem a fusão de informações de fontes heterogêneas.

Este capítulo aborda essa limitação propondo uma arquitetura de RAG Híbrido (HybridRAG) implementada através de um workflow automatizado no N8N unica e exclusivamente para integração ao WhatsApp via Evolution API e que acessa a uma API desenvolvida no projeto em python, a qual implanta o modelo RAGLC. O objetivo é estender as capacidades do Chat para que ele possa, a partir de uma pergunta em linguagem natural, identificar a necessidade de consultar tanto documentos textuais quanto bases de dados estruturadas, recuperar as informações relevantes de ambas as fontes, e sintetizar uma resposta única e coesa para o usuário. Argumentamos que esta capacidade híbrida é um passo fundamental para a criação de assistentes virtuais verdadeiramente úteis e eficientes na administração pública.

4.2 Fundamentos Teóricos: Consultando Dados Estruturados com Linguagem Natural

A ponte entre linguagem natural e dados estruturados é um campo ativo de pesquisa em IA, com abordagens principais:

- **Text-to-SQL:** Converte uma pergunta em linguagem natural diretamente em uma consulta Structured Query Language (SQL) que pode ser executada sobre um banco de dados relacional. LLMs têm mostrado grande capacidade nesta tarefa (ZHU et al., 2024).
- **Knowledge Graph Question Answering (KGQA):** Representa o conhecimento (incluindo dados estruturados) como um grafo de conhecimento (entidades e relações). A pergunta do usuário é mapeada para uma consulta sobre o grafo (ex: SPARQL).
- **RAG sobre Dados Tabulares:** Técnicas para 'linearizar' tabelas de bancos de dados em formato textual, permitindo que sejam indexadas e recuperadas por mecanismos RAG padrão, embora com potencial perda de informação relacional.
- **Agentic Approaches:** Um LLM 'agente' que decide, com base na pergunta, se deve consultar a base vetorial de textos, gerar uma consulta SQL, ou ambos, e depois combina os resultados.

No contexto de HybridRAG, a ideia é integrar uma ou mais dessas técnicas ao pipeline RAG existente. O desafio reside em: a) Roteamento da Consulta: Determinar quais fontes (textual, estruturada) são necessárias para responder. b) Execução da Consulta: Gerar e executar a consulta apropriada para cada fonte (busca vetorial para texto, SQL/SPARQL para estruturado). c) Fusão dos Resultados: Combinar os resultados heterogêneos (trechos de texto, linhas de tabelas) em um formato que o LLM final possa usar para gerar a resposta.

Podemos pensar em uma função de roteamento $Route(Q) \rightarrow \{S_{txt}, S_{sql}, S_{kg}, \dots\}$ que mapeia a pergunta Q para o conjunto de tipos de fontes necessárias. A recuperação $Retrieve(Q, S)$ dependeria do tipo de fonte S . A resposta final A seria gerada a partir da fusão $Merge(\{Retrieve(Q, s) | s \in Route(Q)\})$:

$$A = Generate(Merge(\{Retrieve(Q, s) | s \in Route(Q)\}), Q) \quad (1)$$

4.3 Metodologia Proposta: Arquitetura HybridRAG com N8N

A implementação do sistema HybridRAG foi desenvolvida utilizando uma arquitetura baseada em workflow automatizado através da plataforma N8N, que oferece uma interface visual para criação de fluxos de automação complexos. Esta abordagem permite maior flexibilidade na orquestração dos diferentes componentes do sistema, facilitando a manutenção e evolução da solução.

4.3.1 Arquitetura Geral do Sistema

O sistema é composto por três componentes principais integrados através do workflow N8N:

- ❑ **Interface de Comunicação:** Integração com WhatsApp através da Evolution API, permitindo que usuários interajam com o sistema através de uma interface familiar e amplamente utilizada.
- ❑ **Repositório de Dados:** Google Drive como repositório centralizado para armazenamento tanto dos documentos Portable Document Format (PDF) (editais, leis, portarias) quanto das planilhas Excel contendo dados estruturados sobre candidatos e processos.
- ❑ **Motor de Processamento:** Algoritmo Python RAGLC (RAG com LangChain) para processamento de contextos não estruturados, integrado ao workflow através de chamadas Hypertext Transfer Protocol (HTTP).

4.3.2 Implementação do Workflow N8N

O workflow desenvolvido no N8N, conforme a figura 1, implementa a lógica de roteamento inteligente que constitui o núcleo do sistema HybridRAG. O fluxo operacional é estruturado da seguinte forma:



Figura 1 – Fluxo principal que representa todo o modelo. (Elaborado pelo autor)

Nó de Entrada - Evolution API: O workflow inicia com um webhook que recebe mensagens do WhatsApp através da Evolution API. Este nó é configurado para capturar mensagens de texto dos usuários e extrair informações relevantes como número do remetente, conteúdo da mensagem e timestamp.

Nó de Chamada ao RAGLC: Este nó é o responsável por receber uma *string* com a mensagem enviada pelo usuário. Este texto passará pelo Retrieval Augmented Generation in Logic Context (RAGLC), onde é feito todo o tratamento e o retorno para posterior resposta ao usuário. Mais precisamente, o workflow executa uma chamada HTTP para o algoritmo Python RAGLC, que realiza a busca semântica desenvolvida por este trabalho, processando documentos PDF armazenados no Google Drive ou consultando em uma base de dados de convocações específica e retornando trechos relevantes baseados na similaridade vetorial. A similaridade é apurada pelo próprio algoritmo e a resposta do modelo após o acionamento do LLM é retornada como *string* pela API.

Nó de Resposta - Evolution API: O nó final envia a resposta processada de volta ao usuário através do WhatsApp, utilizando a Evolution API para manter a continuidade da conversa.

4.3.3 Integração com Google Drive

A escolha do Google Drive como repositório central oferece várias vantagens para a implementação do sistema HybridRAG:

Armazenamento Unificado: Tanto os documentos PDF quanto as planilhas Excel são armazenados em uma estrutura de pastas organizada, facilitando o acesso e a manutenção dos dados.

Controle de Versão: O Google Drive mantém automaticamente o histórico de versões dos documentos, permitindo rastreabilidade das alterações e possibilidade de rollback quando necessário.

Acesso Programático: Através das APIs do Google (Drive API e Sheets API), o sistema pode acessar e processar os dados de forma automatizada, sem intervenção manual.

Sincronização Automática: Atualizações nos documentos ou planilhas são imediatamente disponibilizadas para o sistema, garantindo que as respostas sempre reflitam as informações mais recentes.

Segurança e Permissões: O Google Drive oferece controles granulares de acesso, permitindo que apenas usuários autorizados possam modificar os dados, enquanto o sistema mantém acesso de leitura para processamento das consultas.

4.4 Implementação Técnica Detalhada

4.4.1 Configuração da Evolution API

A Evolution API serve como ponte entre o WhatsApp e o workflow N8N. A configuração envolve:

Instância do WhatsApp: Criação de uma instância dedicada na Evolution API, configurada com as credenciais necessárias para envio e recebimento de mensagens.

Webhooks: Configuração de webhooks que direcionam as mensagens recebidas para o endpoint do workflow N8N, garantindo processamento em tempo real.

Autenticação: Implementação de tokens de segurança para garantir que apenas requisições autorizadas sejam processadas pelo sistema.

4.4.2 Algoritmo RAGLC

O algoritmo Python RAGLC, conforme demonstra a figura 2, mantém a funcionalidade core de processamento de documentos não estruturados, com adaptações para integração com o workflow N8N:

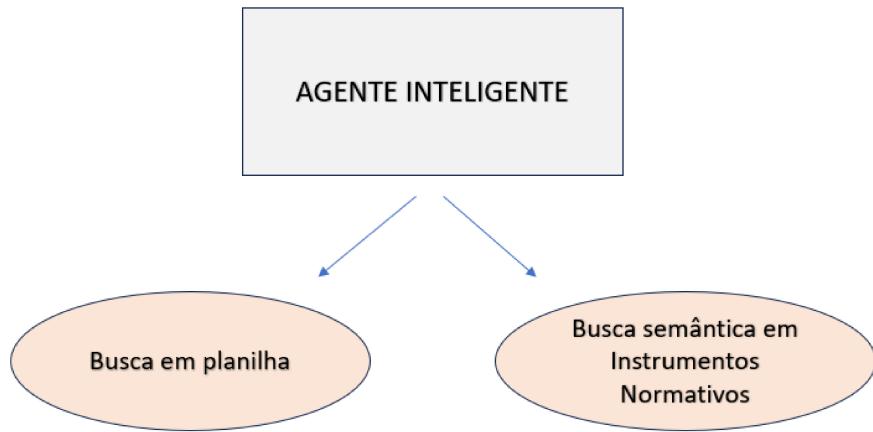


Figura 2 – Representação de funcionalidade do modelo RAGLC. (Elaborado pelo autor)

Interface HTTP: Exposição do algoritmo através de uma API REST, permitindo que o N8N faça chamadas HTTP com a pergunta do usuário e receba a resposta processada.

Acesso ao Google Drive: Integração com a Drive API para acessar dinamicamente os documentos PDF armazenados no repositório, eliminando a necessidade de armazenamento local.

Cache Inteligente: Implementação de sistema de cache para embeddings e índices vetoriais, reduzindo o tempo de resposta para consultas subsequentes.

Processamento Assíncrono: Capacidade de processar múltiplas consultas simultaneamente, melhorando a escalabilidade do sistema.

4.4.3 Lógica de Roteamento Inteligente

O coração do sistema HybridRAG reside na capacidade de determinar automaticamente qual fonte de dados consultar baseado na pergunta do usuário. Esta funcionalidade é implementada através de:

Classificação por Padrões: Identificação de palavras-chave e padrões linguísticos que indicam o tipo de informação solicitada (ex: Cadastro de Pessoa Física (CPF), status, convocação para dados estruturados; requisitos, documentos, procedimentos para dados não estruturados).

Análise Semântica: Utilização de modelos de linguagem para compreender a intenção subjacente da pergunta, mesmo quando não há palavras-chave explícitas.

Decisão Multi-fonte: Capacidade de identificar perguntas que requerem consulta a ambas as fontes, acionando os dois caminhos de processamento em paralelo.

4.5 Fluxo Operacional Detalhado

4.5.1 Cenário 1: Consulta Apenas a Dados Estruturados

Quando um usuário envia uma mensagem como "Qual o status do candidato CPF 123.456.789-00?", o sistema executa o seguinte fluxo:

O nó de entrada recebe a mensagem via Evolution API e extrai o conteúdo textual. O nó de análise identifica a presença do CPF e a palavra "status", classificando a consulta como direcionada a dados estruturados. O workflow aciona o nó de Chamada ao RAGLC, que faz a análise da mensagem e direciona a consulta à planilha para localizar a linha correspondente ao CPF informado. Os dados encontrados (status, data de convocação, etc.) são formatados em linguagem natural. A API retorna a resposta, que é enviada de volta ao usuário através do WhatsApp.

4.5.2 Cenário 2: Consulta Apenas a Documentos Não Estruturados

Para uma pergunta como "Quais documentos são necessários para a posse?", o fluxo é:

O sistema identifica que a pergunta refere-se a procedimentos ou requisitos normativos. O workflow aciona o nó de chamada ao RAGLC, enviando a pergunta via API via HTTP. O algoritmo RAGLC acessa os documentos PDF no Google Drive, executa

a busca semântica e retorna os trechos relevantes. A resposta é processada e enviada ao usuário.

4.5.3 Cenário 3: Consulta Híbrida

Para perguntas complexas como "O candidato CPF 123.456.789-00 foi convocado? Se sim, quais documentos deve levar?", o sistema:

Após receber a mensagem e envio à API, o modelo identifica que a pergunta tem duas partes distintas. Ele aciona simultaneamente a consulta à planilha e aos instrumentos normativos. O modelo apresenta ambos os retornos para o Agente (dentro do algoritmo RAGLC). Uma resposta única e coesa é gerada e enviada ao usuário.

4.6 Vantagens da Arquitetura N8N

A implementação do N8N para integração com o modelo visa facilitar a implementação nos órgãos públicos e oferece várias vantagens práticas e sem custo:

Interface Visual: O workflow pode ser visualizado e modificado através de uma interface gráfica intuitiva, facilitando a manutenção e evolução do sistema.

Modularidade: Cada componente do sistema é representado por nós específicos, permitindo modificações isoladas sem afetar o funcionamento geral.

Escalabilidade: O N8N oferece recursos nativos de escalabilidade, incluindo processamento paralelo e distribuído.

Monitoramento: Ferramentas integradas de logging e monitoramento facilitam a identificação e resolução de problemas.

Integração Nativa: Conectores pré-construídos para APIs REST e webhooks reduzem significativamente o código customizado necessário.

Flexibilidade: Novos nós podem ser facilmente adicionados, caso seja necessária alguma adaptação adicional específica de cada órgão.

4.7 Impacto Potencial e Desafios

HybridRAG transforma o atendimento público aos ingressantes e candidatos:

- ❑ **Respostas Mais Completas e Personalizadas:** Capaz de responder a uma gama muito maior de perguntas, incluindo aquelas específicas sobre a situação individual do usuário (quando autorizado).
- ❑ **Maior Eficiência:** Reduz a necessidade de o usuário consultar múltiplos sistemas ou canais.

- **Novos Serviços:** Abre portas para serviços mais complexos baseados em IA que combinam conhecimento normativo com dados operacionais.
- **Acessibilidade:** A integração com WhatsApp democratiza o acesso às informações, utilizando uma plataforma familiar à maioria dos usuários.

Desafios:

- **Complexidade da Arquitetura:** Integrar múltiplos tipos de consulta e fusão de resultados através de workflows automatizados requer cuidadoso planejamento e teste.
- **Segurança e Lei Geral de Proteção de Dados (LGPD):** Acesso a bases de dados internas através de APIs externas requer cuidados redobrados com segurança, permissões e privacidade.
- **Latência:** Consultas múltiplas e chamadas de API podem aumentar o tempo de resposta, especialmente em cenários híbridos.
- **Dependência de Serviços Externos:** A arquitetura depende da disponibilidade e performance de múltiplos serviços (Google Drive, Evolution API, N8N), requerendo estratégias de contingência.

4.8 Integração com a Dissertação

Esta pesquisa representa um salto significativo em relação ao foco atual da dissertação em RAG puramente textual. Aborda uma limitação fundamental dos sistemas atuais e propõe uma solução com enorme potencial prático para o setor público. A implementação através de workflow N8N demonstra como tecnologias de automação podem ser utilizadas para criar soluções de IA mais acessíveis e mantíveis.

O chatbot proposto pelo trabalho conta com um agente de IA integrado ao WhatsApp via N8N e Evolution, que, ao receber a pergunta do usuário verifica se há correspondência em uma ou mais fontes de dados, direcionando corretamente cada parte da pergunta do usuário para a base correta.

Trata-se de um Agente Especialista em auxiliar ao candidato ou novo admitido, respondendo questões sobre o edital, convocações, benefícios, Plano de Carreira, Direitos e Deveres ou qualquer assunto alusivo ao trabalho constante da base de dados já disponível aos servidores por meio do site oficial, porém, pouco acessada.

A arquitetura proposta demonstra como a combinação de ferramentas modernas de automação (N8N), APIs de comunicação (Evolution API) e repositórios em nuvem (Google Drive) pode criar soluções robustas e escaláveis para o setor público, mantendo baixo custo de implementação e alta eficiência operacional.

4.9 FUNCIONAMENTO DO RAG HÍBRIDO

A ideia central do RAG Híbrido, neste caso, é capacitar um sistema de IA (orquestrado por um workflow N8N) a consultar dinamicamente duas fontes de conhecimento distintas e heterogêneas:

- **Base de Conhecimento Textual (Editais):** Documentos longos, não estruturados ou semi-estruturados (PDFs), ricos em regras, procedimentos e descrições, armazenados no Google Drive. A consulta ideal aqui é a busca por similaridade semântica (RAG tradicional/semântico através do algoritmo RAGLC).
- **Base de Conhecimento Tabular (Planilhas Excel):** Dados organizados em linhas e colunas, com informações específicas, factuais e frequentemente atualizadas (status de candidatos, notas, datas de convocação, contagem de vagas, etc.), também armazenados no Google Drive. A consulta ideal aqui envolve operações de filtragem, seleção, agregação e busca por valores específicos através da Google Sheets API.

4.10 Conclusão

O RAG Híbrido implementado através de workflow N8N representa uma evolução significativa dos sistemas RAG tradicionais, expandindo suas capacidades para além da recuperação de texto e permitindo a integração eficiente com fontes de dados estruturados. Esta abordagem é particularmente valiosa no contexto do serviço público, onde as informações frequentemente estão distribuídas em diferentes formatos e sistemas.

A arquitetura proposta, baseada em workflows visuais que orquestram a comunicação entre diferentes componentes especializados, oferece uma solução mais acessível e mantável para o desafio de integração de dados heterogêneos. A utilização do Google Drive como repositório centralizado, combinada com a integração nativa ao WhatsApp através da Evolution (utilizado no contexto atual apenas para fins acadêmicos), cria um sistema que é tanto tecnicamente robusto quanto acessível aos usuários finais a um investimento significativamente baixo para os moldes do município de Uberlândia ou demais repartições governamentais.

Ao combinar a capacidade de busca semântica em documentos textuais com a precisão das consultas em bancos de dados estruturados, o sistema pode fornecer respostas mais completas, contextualizadas e personalizadas. A implementação através de N8N facilita a manutenção e evolução contínua do sistema, permitindo que administradores sem conhecimento técnico profundo possam fazer ajustes e melhorias no workflow.

Para o contexto específico do atendimento aos candidatos de concursos públicos, esta abordagem permitirá criar um assistente que não apenas esclarece dúvidas sobre o edital, mas também fornece informações personalizadas sobre o status individual do candidato,

combinando o melhor dos dois mundos em uma única interface conversacional acessível através da plataforma de comunicação mais utilizada no Brasil.

Os experimentos e análises apresentados nos próximos capítulos fornecerão evidências sobre a eficácia desta abordagem, validando sua aplicabilidade e identificando oportunidades de aprimoramento contínuo desta arquitetura de RAG Híbrido baseada em workflow automatizado.

CAPÍTULO 5

Experimentos e Análise dos Resultados

Neste capítulo, apresentamos os resultados dos experimentos conduzidos para validar as abordagens propostas neste trabalho. Os experimentos foram organizados para avaliar diferentes aspectos da solução desenvolvida, desde a eficácia dos canais de comunicação até a qualidade das respostas geradas pelo chatbot. Cada seção descreve um conjunto específico de experimentos, seus métodos de avaliação e os resultados obtidos, fornecendo insumos para a validação da hipótese central desta pesquisa.

5.1 Método para a Avaliação

Para avaliar a eficácia da solução proposta, adotamos um conjunto abrangente de metodologias que nos permitiram analisar diferentes dimensões do modelo proposto. Cada aspecto foi avaliado usando métricas e procedimentos específicos, conforme detalhado a seguir:

5.1.1 Avaliação dos Canais de Comunicação

Para a validação dos canais de comunicação, analisamos os tempos de resposta e taxas de comparecimento dos candidatos convocados através de diferentes canais (e-mail e WhatsApp), tendo em vista que a própria Prefeitura de Uberlândia já havia feito um estudo prévio relativo ao telefone. Os dados foram coletados a partir das convocações realizadas no período de abril de 2024, envolvendo candidatos para diferentes cargos.

5.1.2 Avaliação de Modelos de Embeddings e bancos de dados vetoriais

Diferentes modelos de embeddings e bancos de dados vetoriais, populares dentro da comunidade, foram comparados quanto à sua capacidade de representar adequadamente

os textos para recuperação eficiente e performance computacional. Os modelos foram avaliados em termos de percentual de acerto, tempo de execução e consumo de memória.

5.1.3 Avaliação da Segmentação Semântica

A qualidade da segmentação por contexto semântico foi avaliada comparando-a com métodos tradicionais de segmentação por número fixo de tokens. Os resultados foram verificados através de um experimento controlado com 20 perguntas de diferentes níveis de complexidade sobre um texto legal.

5.1.4 Comparação com Respostas corretas ou Oficiais

Para validar a qualidade final do sistema, tanto as respostas geradas pela equipe da área de recrutamento, quanto aquelas produzidas em situações reais (fornecidas pela banca realizadora de processos seletivos), foram comparadas com as respostas emitidas pelo modelo em 6 estruturas testadas. A avaliação considerou uma técnica de Context Recall em um dos casos e em outro foram levantados os pontos em várias métricas com amplo leque de abordagens, conforme detalha o capítulo 3.

5.2 Experimentos

5.2.1 Validação dos Canais de Comunicação

Foram analisados dados de 24 candidatos convocados para diferentes cargos, como Fiscal de Posturas, Iluminador, Agente de Apoio Operacional, Fiscal de Meio Ambiente e Agente de Combate às Endemias. Os principais resultados foram:

- **E-mail:** Das 24 comunicações enviadas por e-mail, nenhuma obteve retorno, resultando em uma taxa de resposta de 0%.
- **WhatsApp:** Das 24 comunicações enviadas por WhatsApp, 19 obtiveram retorno (79,2% de taxa de resposta).
- **Tempo médio de resposta via WhatsApp:** 3 horas e 42 minutos.
- **Taxa de comparecimento:** 70,8% dos candidatos compareceram à convocação.
- **Situações diversas:** 16,7% dos candidatos ainda estavam analisando se assumiriam o cargo, 8,3% remarcaram o atendimento, e 4,2% desistiram formalmente.

Estes resultados demonstram claramente a superioridade do WhatsApp como canal de comunicação em comparação ao e-mail tradicional. Todos os candidatos que compareceram tiveram previamente contato via WhatsApp, evidenciando a importância deste

canal para o sucesso das convocações. Ouvido o Núcleo de Recrutamento sobre a aceitação deste canal de comunicação, foi unânime o retorno de que realmente é o principal e mais efetivo canal de contato com os candidatos. Em relação aos outros meios (telefone e e-mail) O núcleo utiliza uma ferramenta interna para envio de e-mail em massa, para fins de agregação, porém, quase sem efetividade e o telefone não é mais utilizado para esse fim.

5.2.2 Avaliação de Modelos de Embeddings

Foram comparados diferentes modelos de embeddings quanto à sua capacidade de representar adequadamente os textos para recuperação eficiente. A Tabela 2 apresenta um resumo dos resultados. Para esse fim foi proposto um algoritmo de recuperação que mescla um algoritmo genético com vários hiperparâmetros e clusterização com k-means.

Tabela 2 – Comparativo de modelos de embeddings. (Elaborado pelo Autor)

Modelo de embeddings	Acerto	Configuração destacada	Tempo (s)
intfloat/multilingual-e5-large	64,29%	clusters=3, gerações=50, mutação=0,1	1,10
sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2	50,00%	clusters=3, gerações=50, mutação=0,2	0,78
BAAI/bge-large-en-v1.5	50,00%	clusters=3, gerações=5, mutação=0,1	0,52
neuralmind/bert-base-portuguese-cased	35,71%	clusters=3, gerações=50, mutação=0,1	18,0
neuralmind/bert-large-portuguese-cased	21,43%	clusters=3, gerações=100, mutação=0,5	17,1

A comparação entre diferentes modelos de embeddings revelou o intfloat/multilingual-e5-large como o mais eficaz para a tarefa, com 64,29% de acerto perfazendo sua resposta em aproximadamente 6 segundos. Curiosamente, os modelos específicos para o português (neuralmind/bert-base-portuguese-cased e neuralmind/bert-large-portuguese-cased) apresentaram desempenho inferior, apesar de serem treinados especificamente para o idioma dos documentos. Surpreendentemente, os modelos específicos para o português (neuralmind/bert-base-portuguese-cased e neuralmind/bert-large-portuguese-cased) apresentaram desempenho inferior, com 35,71% e 21,43% de acerto, respectivamente.

Em relação ao tempo de execução, todos os modelos tiveram resultados bem semelhantes, tendo performance para a análise comparativa de embeddings para um grupo de 10 perguntas em aproximadamente 0,01 segundos. O modelo intfloat/multilingual-e5-large apenas um dos ciclos apresentou o tempo de 0,02 segundos, porém, como isso ocorreu em somente um dos ciclos e a conexão de internet pode influenciar no processo, esse resultado foi descartado.

. Todos os modelos foram submetidos a vários cenários, todavia, foram testados em condições similares de execução computacional e submetidos aos mesmos parâmetros do Algoritmo Genético (AG).

Além dessa abordagem de medida para os modelos de embedding, o algoritmo também testou de forma eficiente três dos modelos mais populares entre os bancos de dados vetoriais (FAISS, ChromaDB e QDrant), onde o resultado foi um empate técnico entre os modelos que apresentaram tempos de execução médio de 0,01 segundos para verificação em 10 perguntas e cuja acurácia chegou a aproximadamente 78,57% nas iterações realizadas a cada 10 perguntas.

5.2.3 Algoritmo de Segmentação por Contexto Semântico

A seguir é apresentada a avaliação comparativa entre três abordagens de segmentação para o modelo RAG:

- **RAG Tradicional (Naive RAG):** Como ilustrado na 3, Dividindo o texto em chunks de tamanho fixo por número de caracteres, incluindo sobreposição (overlap) para ampliar o contexto.

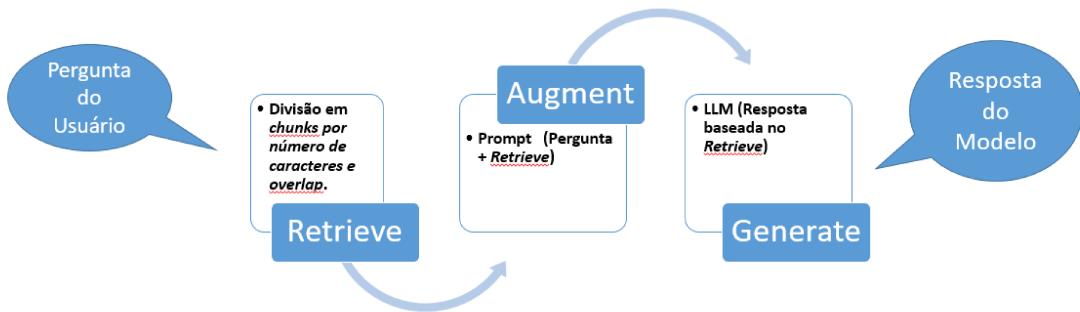


Figura 3 – Fluxograma do RAG Tradicional (Naive RAG) com divisão de chunks por tamanho fixo. (Elaborado pelo autor)

- **RAG por Bloco:** A figura 4 apresenta o fluxo completo, onde cada bloco representa um artigo completo, incluindo incisos e parágrafos, respeitando a estrutura original do documento legal, ou, no caso de artigos são quebrados em parágrafos (em sua maioria subitens dentro de um edital).
- **RAG Semântico (RAGLC):** Nesta abordagem (figura 5), O texto é dividido em blocos, assim como no caso anterior, porém, conta com mais uma iteração responsável por agrupar blocos semelhantes no contexto semântico, medida na transformação de cada bloco em embeddings e comparando os embeddings entre si por distância de cosseno. Os chunks mais semelhantes são agrupados com base nos parâmetros descritos no capítulo 3. Nesse contexto foi criada ainda mais uma camada de agrupamento semântico utilizando clusterização semântica.

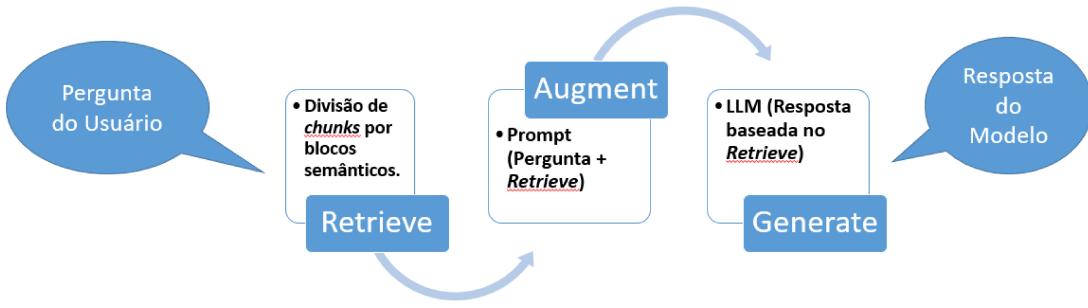


Figura 4 – Fluxograma do RAG por Bloco, com divisão baseada em unidades semânticas do documento. (Elaborado pelo autor)

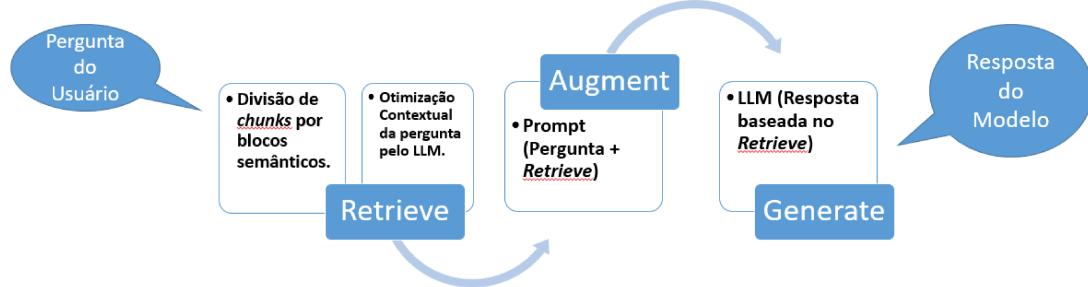


Figura 5 – Fluxograma do RAG Semântico (RAGLC) proposto, que adiciona uma etapa de otimização contextual da pergunta via LLM. (Elaborado pelo autor)

5.2.3.1 Primeiro Teste (Lei Municipal com Context Recall)

Para este primeiro experimento a Tabela 4 apresenta os resultados resumidos comparando as três abordagens, onde foi usado como base de dados a Lei Complementar Municipal nº 40/1992 (Estatuto dos Servidores Públicos Municipais de Uberlândia) e a Tabela 5, apresenta os resultados detalhados por pergunta para cada modelo. Foram utilizadas 20 perguntas extraídas da Lei Complementar nº 40/1992.

Os resultados mostram que a abordagem de RAG Semântico superou significativamente as outras abordagens, especialmente nas perguntas mais complexas (11-15 e 16-20) que exigiam a interpretação de múltiplos artigos ou blocos maiores de conteúdo. A acurácia geral do modelo semântico foi de 98%, contra 93% do RAG Tradicional e 92% do RAG por Artigo.

Tabela 3 – Lista de Perguntas sobre o Estatuto dos Servidores. (Elaborado pelo Autor)

Perguntas
1. O que é considerado servidor público segundo a Lei Complementar nº 40/1992?
2. Como a lei define “cargo público”?
3. De que forma os cargos de provimento efetivo são organizados?
4. O que caracteriza uma classe de cargos, segundo a lei?
5. Quais são os requisitos básicos para ingresso no serviço público?
6. Como ocorre a investidura em cargo público?
7. Quais são as formas de provimento em cargo público previstas na lei?
8. Qual é o prazo de validade do concurso público, de acordo com a lei?
9. Em quanto tempo o servidor adquire estabilidade no serviço público?
10. Quais são as causas que configuram a vacância do cargo público?
11. Quais os procedimentos e prazos para a posse do servidor, considerando os dispositivos sobre a assinatura do termo e início do exercício?
12. De que forma o início do exercício do cargo é regulamentado e como deve ser feito o registro das mudanças?
13. Como a promoção é definida e qual seu critério para progressão na carreira?
14. Quais condições limitam a reversão do servidor aposentado?
15. Quais os critérios para a redistribuição de servidores entre órgãos ou entidades?
16. De que forma a lei protege o servidor estável quanto à perda do cargo?
17. Como é realizada a contagem do tempo de serviço para efeitos de aposentadoria e quais elementos são considerados?
18. Qual a importância do concurso público para o provimento dos cargos efetivos e quais os principais critérios exigidos?
19. De que maneira a lei assegura o direito de pessoas portadoras de deficiência de participar dos concursos públicos?
20. Quais são as modalidades de exoneração do servidor e as condições que as regulam?

Tabela 4 – Resultado por blocos de perguntas de modelos RAG

Algoritmo	Perguntas 1-10	Perguntas 11-15	Perguntas 16-20	Acurácia
RAG Tradicional	9,8	8,5	8,9	93%
RAG por Artigo	9,9	8,3	8,6	92%
RAG Semântico	10,0	9,6	9,9	98%

Tabela 5 – Resultados de pontuação (0–10) para cada uma das 20 perguntas.

Pergunta	RAG Tradicional	RAG por Artigo	RAG Semântico
1	10,00	10,00	10,00
2	10,00	10,00	10,00
3	10,00	9,00	9,50
4	10,00	10,00	10,00
5	9,00	10,00	10,00
6	10,00	10,00	10,00
7	10,00	10,00	10,00
8	10,00	10,00	10,00
9	10,00	10,00	10,00
10	10,00	10,00	10,00
11	8,50	5,00	7,50
12	8,50	9,00	10,00
13	10,00	10,00	10,00
14	6,00	7,50	9,50
15	9,00	10,00	10,00
16	10,00	10,00	10,00
17	9,50	10,00	10,00
18	10,00	5,00	10,00
19	10,00	10,00	10,00
20	5,00	8,00	10,00
Acurácia	93%	92%	98%

5.2.3.2 Segundo Teste (Edital testado em 6 cenários)

Algoritmo 1 SemanticChunker

INPUT: Documento D, Modelo_EMBEDDING E, Limiar_Similaridade T

OUTPUT: Lista de Chunks C

1. Blocos_Iniciais B = Segmentar D por estrutura
 2. Embeddings_Blocos Emb = Gerar embeddings para cada bloco
 3. Matriz_Similaridade S = Calcular similaridade_cosseno(Emb)
 4. Grafo G = Criar grafo com nos = {0..|B|-1}
 5. PARA cada par (i, j) com i < j:
 6. SE S[i, j] >= T:
 7. Adicionar aresta (i, j) em G
 8. Comunidades Com = Encontrar_componentes_conexas(G)
 9. Chunks C = []
 10. PARA cada comunidade em Com:
 11. IDs_Oordenados = Ordenar IDs da comunidade
 12. Chunk_Atual = [B[i] para i em IDs_Oordenados]
 13. Adicionar Chunk_Atual em C
 14. RETORNAR C
-

Em uma segunda abordagem, o modelo foi avaliado comparando as respostas oficiais fornecidas pela banca examinadora do Processo Seletivo nº 01/2025. Foi ainda incluída

uma abordagem para clusterização semântica (Algoritmo 1), conforme detalhado a seguir.

Abaixo são apresentadas as etapas de execução do algoritmo:

- ❑ **Geração de Embeddings:** $O(N \cdot L)$, onde N é o número de blocos e L é o custo de processar um bloco.
- ❑ **Cálculo da Matriz de Similaridade:** $O(N^2 \cdot D)$, onde D é a dimensão dos embeddings.
- ❑ **Construção do Grafo e Detecção de Comunidades:** Eficientemente implementável em $O(N^2)$.
- ❑ **Complexidade Geral:** Dominada pelo cálculo da matriz de similaridade, resultando em $O(N^2)$.

Para determinar a arquitetura de RAG mais eficaz, foi conduzida uma análise comparativa, expandindo o escopo experimental para avaliar seis estratégias distintas. Estas estratégias resultam da combinação de três técnicas de segmentação de texto (chunking) com duas abordagens de formulação de consulta (query). O objetivo foi mensurar não apenas o impacto de cada componente isoladamente, mas também as combinações entre eles.

As seis configurações avaliadas foram:

1. **RAG Tradicional + Consulta Original (`rag_orig`):** Utiliza segmentação por tamanho fixo com sobreposição e a pergunta original do usuário.
2. **RAG Tradicional + Consulta Enriquecida (`rag_enr`):** Utiliza segmentação por tamanho fixo, mas a consulta é pré-processada e expandida por um LLM.
3. **RAG por Contexto Simples + Consulta Original (`ctx_orig`):** Emprega uma segmentação baseada no agrupamento sequencial de blocos de texto semanticamente coesos do documento original, utilizando a pergunta original.
4. **RAG por Contexto Simples + Consulta Enriquecida (`ctx_enr`):** Combina o agrupamento sequencial de blocos com a consulta enriquecida.
5. **RAG por Contexto Complexo + Consulta Original (`ctxCX_orig`):** Aplica a técnica de clusterização semântica, que agrupa blocos de informação com base na similaridade de embeddings, utilizando a pergunta original.
6. **RAG por Contexto Complexo + Consulta Enriquecida (`ctxCX_enr`):** Combina a clusterização semântica com a consulta enriquecida, representando a abordagem mais sofisticada.

O desempenho de cada estratégia foi avaliado por um conjunto de métricas, culminando em um score ponderado geral que sintetiza a qualidade da resposta. Conforme ilustrado na Figura 6, a estratégia `ctxCX_enr` alcançou o maior desempenho, validando a hipótese de que a combinação de técnicas avançadas de segmentação e enriquecimento de consulta produz os melhores resultados.

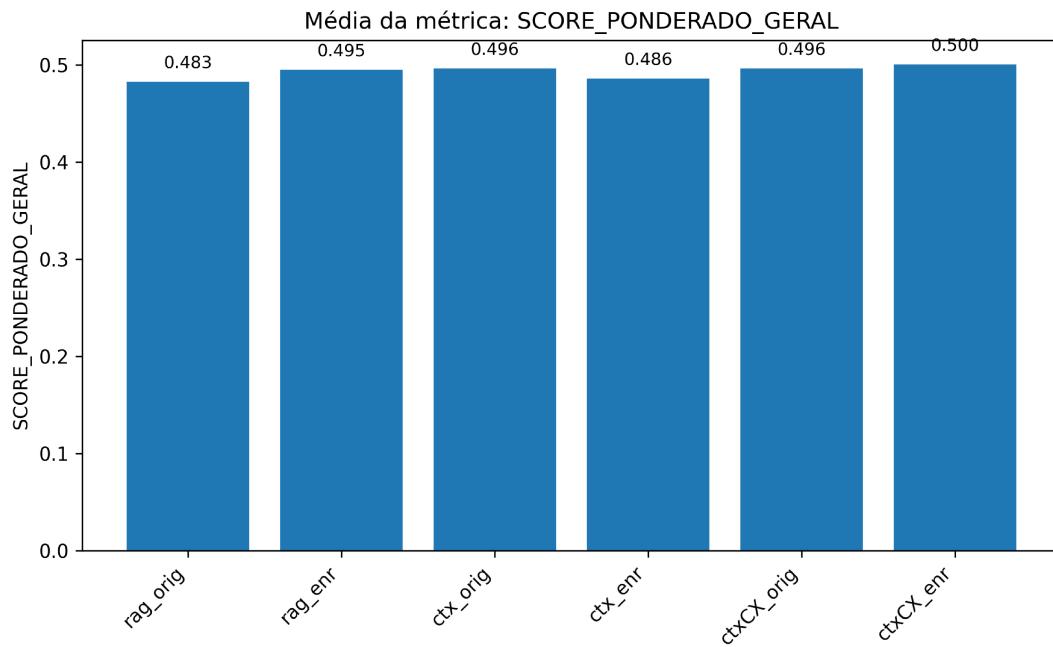


Figura 6 – Comparando as seis estratégias de RAG. (Elaborado pelo Autor)

Análise dos Resultados

A análise dos scores revela nuances importantes sobre a interação entre os componentes do pipeline RAG.

Impacto da Estratégia de Segmentação

Ao comparar os modelos com a mesma estratégia de consulta (e.g., `rag_orig` vs. `ctx_orig` vs. `ctxCX_orig`), observa-se uma clara superioridade das abordagens baseadas em contexto. Os modelos `ctx_orig` (0.496) e `ctxCX_orig` (0.496) superaram significativamente o RAG tradicional `rag_orig` (0.483). Isso demonstra que preservar a coerência semântica dos blocos de informação durante a segmentação é mais eficaz do que a divisão por tamanho fixo. A abordagem `ctxCX_enr`, que utiliza clusterização semântica, obteve o score máximo (0.500), confirmando que a criação de chunks tematicamente coesos, mesmo que não sequenciais no texto original, fornece o contexto mais relevante para o LLM.

Impacto do Enriquecimento de Consulta

O efeito do enriquecimento de consulta não foi uniformemente positivo, revelando uma forte dependência da técnica de segmentação subjacente.

- No **RAG Tradicional**, o enriquecimento gerou uma melhora notável, elevando o score de 0.483 (`rag_orig`) para 0.495 (`rag_enr`).
- No **RAG por Contexto Complexo**, a melhoria foi mais sutil, mas decisiva, levando ao melhor desempenho geral: de 0.496 (`ctxCX_orig`) para 0.500 (`ctxCX_enr`).
- Surpreendentemente, no **RAG por Contexto Simples**, o enriquecimento foi prejudicial, causando uma queda no desempenho de 0.496 (`ctx_orig`) para 0.486 (`ctx_enr`). Uma possível explicação é que a consulta enriquecida, ao se tornar mais específica, pode ter limitado a busca a um escopo que não se alinhava bem com a estrutura de blocos sequenciais, que é mais ampla.

Análise de Métricas Específicas

A análise das métricas individuais oferece insights complementares.

- **Métricas Semânticas (BERTSCORE, COSINE_ST)**: Todas as estratégias produziram respostas com alta similaridade semântica em relação à referência, com os modelos `ctxCX` (`ctxCX_orig` com 0.675 no BERTSCORE e `ctxCX_enr` com 0.896 no COSINE_ST) apresentando uma leve vantagem, indicando maior fidelidade contextual.
- **Métricas Lexicais (F1_OVERLAP, JACCARD, BLEU)**: A estratégia `ctx_orig` se destacou consistentemente nessas métricas (F1_OVERLAP: 0.373, JACCARD: 0.236, BLEU: 0.077). Isso sugere que sua abordagem de agrupar blocos sequenciais tende a preservar a formulação e a densidade de palavras-chave mais próximas das respostas de referência.
- **Similaridade TF-IDF (COSINE_TFIDF)**: Os modelos com consulta enriquecida, especialmente `ctxCX_enr` (0.461), demonstraram superioridade, o que é esperado, já que o enriquecimento tende a focar a busca em termos específicos e de alta relevância, que são valorizados pelo TF-IDF.

Os resultados validam a hipótese central (H1) da pesquisa. A combinação da **segmentação semântica** com o **enriquecimento de consulta** provou ser a estratégia mais eficaz, superando o baseline (Naive RAG com Query Original) em todas as métricas agregadas.

A superioridade da segmentação semântica pode ser atribuída à sua capacidade de criar *chunks* de contexto mais coesos e completos, evitando a fragmentação de informações cruciais que ocorre na divisão por tamanho fixo. O enriquecimento da consulta, por sua vez, ajuda a alinhar a intenção do usuário com o vocabulário técnico do edital, melhorando a precisão da recuperação vetorial.

Relevância das diferenças observadas

Embora, em alguns casos, as variações quantitativas entre os resultados possam parecer pequenas à primeira vista, sua relevância é significativa no contexto desta pesquisa. Isso se deve a três fatores principais:

1. **Multiplicidade de métricas:** A avaliação foi conduzida a partir de um conjunto diversificado de métricas (semânticas, lexicais e baseadas em recuperação). Quando um modelo apresenta ganhos consistentes, ainda que sutis, em diferentes métricas, o efeito combinado se traduz em uma melhora substancial na qualidade da resposta final.
2. **Impacto acumulativo:** Pequenas diferenças em métricas como *recall*, *precision* ou *cosine similarity* podem alterar completamente o conjunto de *chunks* recuperados. Em um cenário real de atendimento ao candidato, isso pode significar a diferença entre uma resposta completa e contextualizada ou uma resposta parcial/inadequada. Assim, variações aparentemente discretas podem gerar consequências práticas bastante distintas.
3. **Natureza não linear dos LLMs:** Os modelos de linguagem tendem a amplificar diferenças de entrada. Pequenas mudanças no contexto recuperado influenciam diretamente o processo de geração da resposta, resultando em saídas que divergem de forma significativa em termos de correção e clareza. Portanto, pequenas diferenças nos experimentos refletem impactos desproporcionais na experiência final do usuário.

Em síntese, as diferenças numéricas observadas não devem ser interpretadas como pequenas. Pelo contrário, quando analisadas em conjunto e sob múltiplas perspectivas de avaliação, revelam-se determinantes para a comprovação da hipótese central desta pesquisa e para a validação da eficácia da abordagem proposta.

- **Caso de Sucesso (Pergunta sobre inscrição em dois cargos):** A pergunta “Posso me inscrever para dois cargos?” exige a correlação de múltiplos itens do edital. O RAG Semântico foi capaz de recuperar e sintetizar todos os itens relevantes (3.6, 3.6.1, 3.6.2), fornecendo uma resposta completa. O RAG Tradicional, por sua vez, recuperou apenas um dos itens, resultando em uma resposta parcialmente correta, mas incompleta (7).

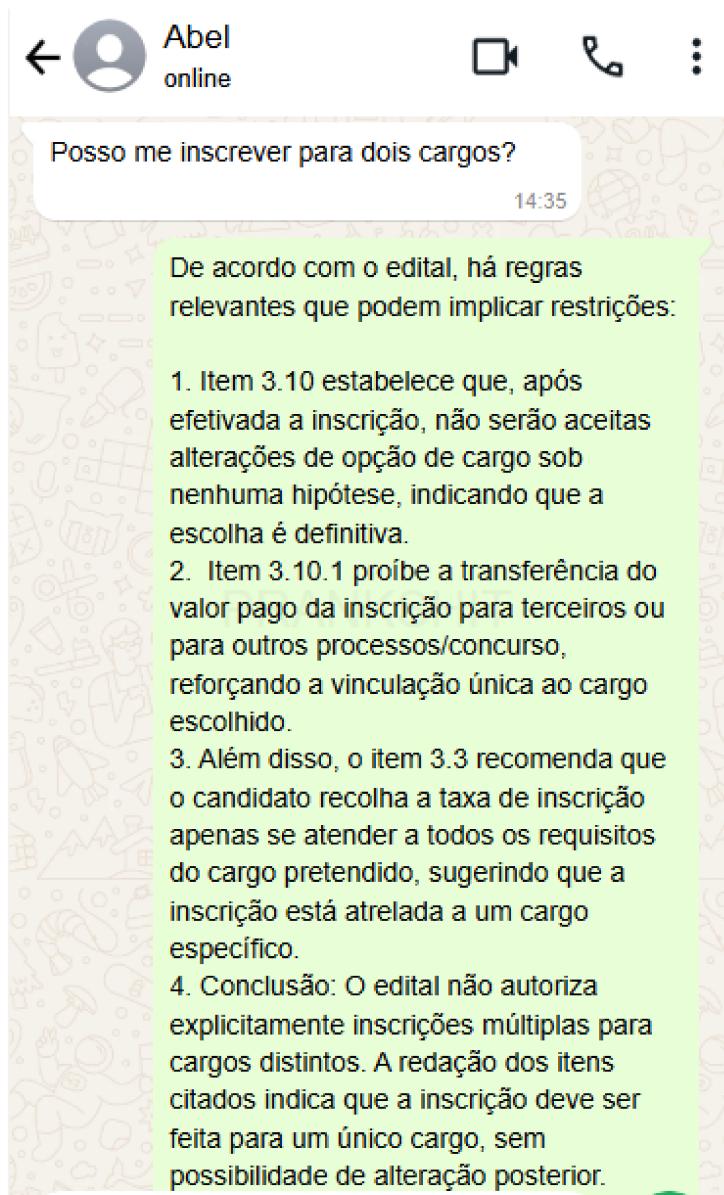


Figura 7 – Resposta bem sucedida do modelo. Contexto recuperado corretamente. (Elaborado pelo Autor)

□ **Padrão de Erro (Pergunta sobre local de trabalho):** Para a pergunta “Passei no concurso e preciso saber onde vou trabalhar”, a maioria dos modelos teve dificuldade. A resposta correta (“os locais de trabalho serão definidos no ato da contratação”) está em um item curto (1.2). O RAG Tradicional tendeu a recuperar *chunks* maiores sobre a etapa de contratação, que não continham a resposta específica, levando à resposta “Não foram encontradas informações”. Isso evidencia uma limitação onde *chunks* maiores podem “afogar” a informação específica e curta, porém, nesse contexto o RAGLC também não foi capaz de recuperar a resposta correta (8).

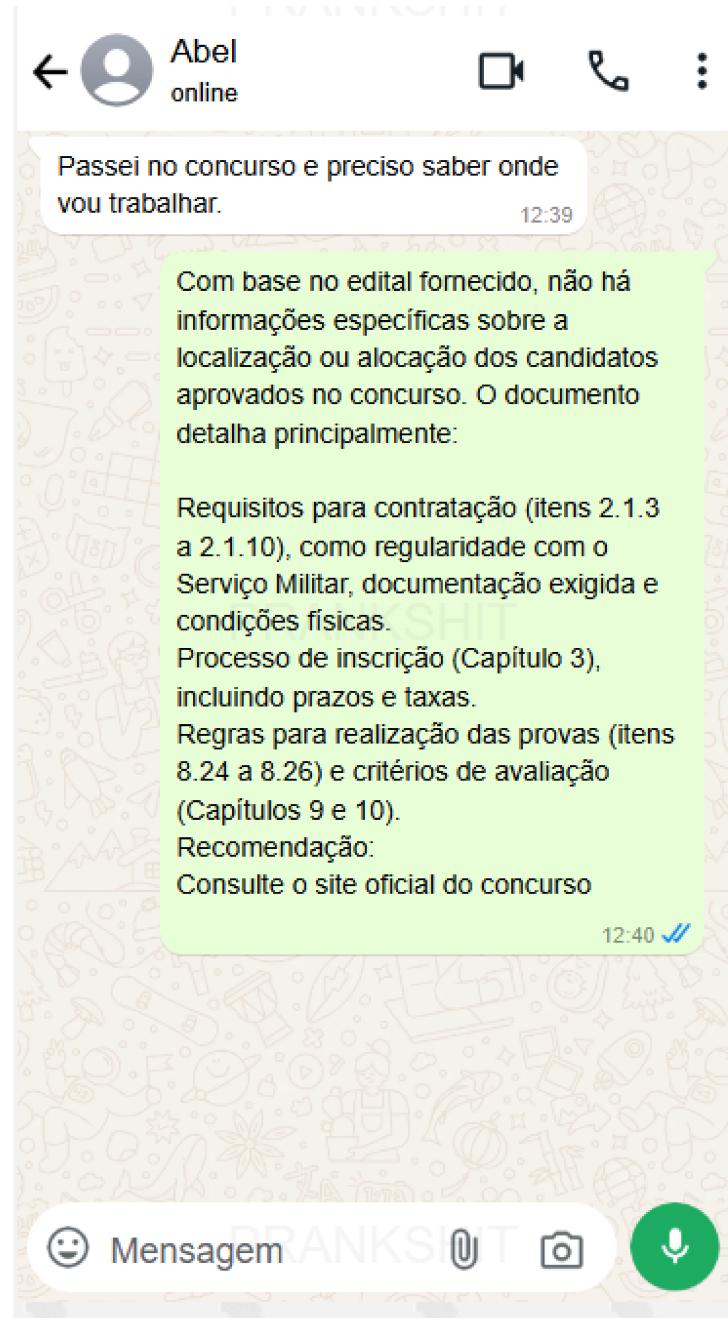


Figura 8 – Resposta do Agente que não conseguiu recuperar o contexto correto. (Elaborado pelo Autor)

5.2.4 Avaliação Geral

Considerando os resultados em conjunto, podemos afirmar que a hipótese central desta pesquisa foi validada: a utilização de técnicas avançadas de NLP e LLMs efetivamente melhora a eficiência, a qualidade das respostas e a celeridade do atendimento ao cidadão no contexto abordado por este trabalho.

Especificamente, os experimentos demonstraram:

- ❑ A superioridade do WhatsApp como canal de comunicação (79,2% de taxa de res-

posta).

- A identificação do modelo de embedding mais adequado (intfloat/multilingual-e5-large, com 64,29% de acerto).
- Um empate técnico em os bancos de dados vetoriais mais utilizados na atualidade.
- A clara vantagem da segmentação semântica sobre métodos tradicionais (98% vs. 93% de acurácia).
- O melhoramento significativo da contextualização quando utilizado o conjunto ***query enrichment + segmentação por contexto +***.
- A capacidade do chatbot de fornecer respostas mais completas, personalizadas e rápidas que o sistema tradicional de atendimento.

As limitações observadas, como a dificuldade de recuperação de todos os modelos em perguntas de tipos e formatos específicos, representam oportunidades de melhoria para trabalhos futuros, possivelmente através de técnicas mais avançadas de verificação factual ou mecanismos de feedback humano.

Em conclusão, os resultados experimentais fornecem evidências robustas de que a abordagem proposta pode efetivamente transformar o atendimento ao ingressante e ao candidato de processos seletivos e concursos públicos, combinando a precisão e confiabilidade dos procedimentos oficiais com a eficiência, personalização e acessibilidade proporcionadas pelas tecnologias de IA.

CAPÍTULO 6

Conclusão

Este trabalho abordou o desenvolvimento de um chatbot baseado em técnicas avançadas de PLN e LLMs para automação do atendimento a ingressantes e candidatos de processos seletivos na Prefeitura Municipal de Uberlândia. O objetivo foi mitigar a sobrecarga de trabalho dos servidores públicos e melhorar a qualidade do atendimento ao cidadão, através de um sistema capaz de processar e responder a consultas em linguagem natural de forma eficiente e contextualizada.

Partindo da hipótese de que modelos de PLN e LLMs poderiam melhorar a eficiência, qualidade e celeridade do atendimento público, conduzimos uma série de experimentos que não apenas validaram essa premissa, mas revelaram insights valiosos sobre as melhores abordagens para implementação desses sistemas no contexto da administração pública.

A pesquisa envolveu múltiplas dimensões de análise, desde a avaliação dos canais de comunicação mais eficazes até a comparação detalhada de diferentes modelos de embeddings e técnicas de segmentação semântica. Os resultados obtidos demonstraram claramente que a abordagem proposta superou os métodos tradicionais em aspectos críticos como completude, personalização e qualidade de resposta, oferecendo uma alternativa viável, de baixo custo e eficiente aos métodos convencionais de atendimento.

6.1 Principais Contribuições

Este trabalho apresenta cinco contribuições principais para o avanço da aplicação de tecnologias de IA na administração pública:

- **Desenvolvimento de um modelo de atendimento automatizado para processos seletivos públicos:** Implementação um chatbot especializado no domínio de informações relevantes aos servidores municipais e a candidatos de concursos públicos e processos seletivos, capaz de interpretar e responder a perguntas complexas sobre esses assuntos.

- ❑ **Avaliação comparativa de modelos de embeddings para o domínio público:** A análise detalhada de cinco modelos de embeddings (intfloat/multilingual-e5-large, sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2, BAAI/bge-large-en-v1.5, neuralmind/bert-base-portuguese-cased, neuralmind/bert-large-portuguese) forneceu evidências empíricas sobre seu desempenho no contexto específico de documentos públicos em português, identificando o intfloat/multilingual-e5-large como o mais adequado para esta aplicação, com 64,29% de acerto.
- ❑ **Desenvolvimento de uma técnica de segmentação semântica para documentos legais:** A abordagem de segmentação por blocos semânticos, em vez da divisão tradicional por tokens, demonstrou superioridade significativa, alcançando 98% de acurácia contra 93% da abordagem tradicional. Esta técnica preserva a integridade semântica dos documentos legais e melhora a qualidade das respostas, especialmente para consultas complexas que exigem a interpretação de múltiplos trechos do documento. Além disso foi possível observar que técnicas de *query enrichment* melhoraram a qualidade do *Retrieve* e consequentemente a qualidade das respostas do RAGLC.
- ❑ **Proposta de arquitetura RAG Híbrido:** Desenvolvimento de uma arquitetura inovadora de RAG Híbrido que integra dados estruturados e não estruturados, permitindo ao chatbot consultar documentos textuais quanto bases de dados operacionais. Esta abordagem amplia o escopo das perguntas que podem ser respondidas automaticamente, combinando informações do edital com dados específicos sobre candidatos e processos.

Além das contribuições técnicas e metodológicas, o trabalho também demonstrou o potencial transformador da aplicação de tecnologias de IA no âmbito da Administração Pública. Os experimentos revelaram que o chatbot proposto pode atender às necessidades de mais de aproximadamente 20.000 candidatos anuais com qualidade superior e tempo de resposta drasticamente reduzido, gerando economia de recursos e melhorando a experiência do cidadão.

6.2 Trabalhos Futuros

Apesar dos resultados promissores, identificamos várias oportunidades de aprimoramento e expansão da pesquisa:

- ❑ **Aperfeiçoamento da precisão factual:** Embora o chatbot tenha alcançado boa precisão em um ambiente produzido, ainda há uma margem para melhoria em relação às respostas oficiais (ambiente realístico). Trabalhos futuros podem explorar técnicas de verificação factual e mecanismos de feedback humano para reduzir esta

diferença, possivelmente através de abordagens de (Aprendizado por Reforço com Feedback Humano (RLHF)).

- **Expansão da arquitetura RAG Híbrido:** A integração entre dados estruturados e não estruturados pode ser refinada e expandida para incorporar outras fontes de dados, como sistemas legados, bancos de dados relacionais complexos e fontes externas. Pesquisas futuras podem explorar mecanismos mais sofisticados de roteamento de consultas e fusão de resultados para criar um sistema verdadeiramente abrangente.
- **Multilinguagem e acessibilidade:** Desenvolver capacidades multilíngues e recursos de acessibilidade, como integração com interfaces de voz, para atender a uma população mais diversificada.
- **Implementação de técnicas de explicabilidade:** Incorporar mecanismos que permitam ao chatbot explicar seu raciocínio e fontes de informação, aumentando a transparência e a confiança nas respostas fornecidas.
- **Adaptação para outros domínios públicos:** Aplicar a metodologia desenvolvida a outros domínios da administração pública, como saúde, educação e assistência social, e outros órgãos públicos de todas as esferas, adaptando os modelos e técnicas às especificidades de cada área.
- **Análise de impacto econômico e social:** Realizar estudos quantitativos sobre o impacto econômico (economia de recursos) e social (satisfação do cidadão, inclusão digital) da implementação em larga escala de chatbots na administração pública.

A segmentação semântica, em particular, mostrou-se uma área promissora para pesquisas futuras. Técnicas mais avançadas de detecção de estruturas semânticas em documentos complexos, possivelmente utilizando modelos especializados ou técnicas de zero-shot learning, poderiam melhorar ainda mais a qualidade da recuperação de contexto.

6.3 Contribuições em Produção Bibliográfica

Durante o desenvolvimento desta pesquisa, foi produzida a contribuição bibliográfica abaixo, submetida na conferência:

- BORGES, A. M.; SOUZA, J. R.; ALVES, R. C. M. F. "RAGLC: A Semantic Block-Based RAG Model for Legal Document Contextual Retrieval" In: Revista de Informática Teórica e Aplicada, 2025. (Submetido)

Além da submissão para publicação acadêmica, o projeto resultou no desenvolvimento de um modelo estruturado e de baixo custo que será apresentado à Prefeitura Municipal

de Uberlândia, com potencial para beneficiar diretamente os mais de 50.000 candidatos e servidores públicos em geral, reduzindo significativamente a carga de trabalho de diversos setores e em especial aos servidores do Núcleo de Recrutamento, ligado à Secretaria Municipal de Administração.

Em conclusão, este trabalho demonstrou o potencial transformador da aplicação de técnicas avançadas de PLN e LLMs no contexto da Administração Pública. Os resultados alcançados não apenas validam a hipótese inicial de que essas tecnologias podem melhorar significativamente a eficiência e qualidade do atendimento ao cidadão, mas também abrem caminho para uma nova geração de serviços públicos mais inteligentes, eficazes, responsivos e centrados no cidadão.

Referências

- BRASIL. O que é o portal gov.br e como o cidadão participa do gov.br? 2021. <Https://www.gov.br/pt-br/sobre#: :text=O>cesso em: 7 fev. 2024.
- _____. Quanto essa unificação vai demandar de investimento do Governo Federal? 2021. <Https://www.gov.br/pt-br/sobre#: :text=O>cesso em: 7 fev. 2024.
- _____. Central de Atendimento lança chatbot para comunicação com usuários dos sistemas da Gestão. 2024. <Https://www.gov.br/gestao/pt-br/assuntos/noticias/2024/novembro/central-de-atendimento-lanca-chatbot-para-comunicacao-com-usuarios-dos-sistemas-da-gestao>. Acesso em: 3 dez. 2024.
- _____. Inteligência Artificial na Administração Pública. 2025. <Https://revistaft.com.br/inteligencia-artificial-na-administracao-publica/>. Acesso em: 27 jan. 2025.
- CHANDEL, S.; SHUKLA, A.; GUPTA, A. Chatbot: Efficient and utility-based platform. **SAI 2018**, Springer Nature Switzerland AG, p. 109–122, 2018.
- CHO, K. et al. Learning phrase representations using rnn encoder-decoder for statistical machine translation. **arXiv preprint arXiv:1406.1078**, 2014.
- CHOI, F.; WIEMER-HASTINGS, P.; MOORE, J. Automatic topic segmentation and labeling for long documents. In: **Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics**. [S.l.: s.n.], 2000. p. 707–714.
- DeepSeek AI. **DeepSeek API Documentation**. 2024. Acesso em: 26 ago. 2025. Disponível em: <<https://api-docs.deepseek.com/>>.
- _____. **DeepSeek API Documentation – Token Usage**. 2024. Acesso em: 26 ago. 2025. Disponível em: <https://api-docs.deepseek.com/quick_start/token_usage>.
- _____. **DeepSeek API – Create Chat Completion**. 2024. Acesso em: 26 ago. 2025. Disponível em: <<https://api-docs.deepseek.com/api/create-chat-completion>>.
- _____. **DeepSeek-V3 Technical Report**. 2024. ArXiv preprint. Acesso em: 26 ago. 2025. Disponível em: <<https://arxiv.org/abs/2412.19437>>.
- DEVLIN, J. et al. **BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding**. 2019.

- ENGIN, Z.; TRELEAVEN, P. Algorithmic government: Automating public services and supporting civil servants in using data science technologies. **The Computer Journal**, v. 62, n. 3, p. 449–461, 2018. Open Access under CC BY 4.0.
- ES, S. et al. **Ragas: Automated Evaluation of Retrieval Augmented Generation**. 2025. Disponível em: <<https://arxiv.org/abs/2309.15217>>.
- GAO, L. et al. Precise zero-shot dense retrieval without relevance labels. **Transactions of the Association for Computational Linguistics**, v. 11, p. 624–638, 2023.
- GAO, Y. et al. **Retrieval-Augmented Generation for Large Language Models: A Survey**. 2024. Disponível em: <<https://arxiv.org/abs/2312.10997>>.
- HEARST, M. A. Texttiling: Segmenting text into multi-paragraph subtopic passages. In: MIT PRESS. **Computational Linguistics**. [S.l.], 1997. v. 23, n. 1, p. 33–64.
- HOCHREITER, S.; SCHMIDHUBER, J. Long short-term memory. **Neural Computation**, MIT Press, v. 9, n. 8, p. 1735–1780, 1997.
- HOLLAND, J. H. Genetic algorithms and the optimal allocation of trials. **SIAM Journal on Computing**, SIAM, v. 2, n. 2, p. 88–105, 1973.
- HUANG, K. et al. **CRISPR-GPT: An LLM Agent for Automated Design of Gene-Editing Experiments**. 2024. Disponível em: <<https://arxiv.org/abs/2404.18021>>.
- KLEIMAN, F.; BARBOSA, M. M. **Management and Performance Program Chatbot: A Use Case of Large Language Model in the Federal Public Sector in Brazil**. [S.l.]: ACM, 2024.
- LEWIS, P. et al. **Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks**. 2020.
- _____. Retrieval-augmented generation for knowledge-intensive nlp tasks. **Advances in Neural Information Processing Systems**, v. 33, p. 9459–9474, 2020.
- _____. **Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks**. 2021. Disponível em: <<https://arxiv.org/abs/2005.11401>>.
- LIN, C.-Y. Rouge: A package for automatic evaluation of summaries. In: **Text Summarization Branches Out: Proceedings of the ACL-04 Workshop**. [S.l.: s.n.], 2004. p. 74–81.
- LIU, D. et al. **DeepSeek LLM: Scaling Open-Source Language Models with Longtermism**. 2024. Disponível em: <<https://arxiv.org/abs/2412.19437v1>>.
- MCCARTHY, J. et al. A proposal for the dartmouth summer research project on artificial intelligence. **AI Magazine**, AAAI Press, v. 27, n. 4, p. 12–14, 2006.
- OSBORNE, S.; BROWN, K. **Managing change and innovation in public service organizations**. Oxon: Routledge, 2005.
- PAIVA, H.; FERNANDES, A. L.; ALTURAS, B. **IT Outsourcing Success Factors in the Portuguese Public Sector: A research proposal**. Madrid, Espanha: [s.n.], 2022. 1–6 p.

- PAPINENI, K. et al. Bleu: a method for automatic evaluation of machine translation. In: **Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics**. [S.l.: s.n.], 2002. p. 311–318.
- RADFORD, A. et al. Improving language understanding by generative pre-training. **OpenAI Blog**, v. 1, n. 8, p. 1–12, 2018.
- RAFFEL, C. et al. Exploring the limits of transfer learning with a unified text-to-text transformer. **Journal of Machine Learning Research**, v. 21, n. 140, p. 1–67, 2020.
- REIMERS, N.; GUREVYCH, I. Sentence-bert: Sentence embeddings using siamese bert-networks. In: **Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing**. [S.l.: s.n.], 2019.
- _____. **Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks**. 2019. 3973–3983 p.
- _____. Sentence-bert: Sentence embeddings using siamese bert-networks. In: **Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing**. Association for Computational Linguistics, 2019. Disponível em: <<http://arxiv.org/abs/1908.10084>>.
- RUSSELL, S.; NORVIG, P. **Artificial Intelligence: A Modern Approach**. 4. ed. [S.l.]: Pearson, 2020. ISBN 9780134610993.
- RUSSELL, S. J.; NORVIG, P. **Artificial Intelligence: A Modern Approach**. [S.l.]: Pearson Education, Limited, 2020.
- SCHUMPETER, J. A. **The theory of economic development: an inquiry into profits, capital, credit, interest, and the business cycle**. Nova Brunsique: Transaction Publishers, 1934.
- SOUZA, F.; NOGUEIRA, R.; LOTUFO, R. BERTimbau: pretrained BERT models for Brazilian Portuguese. In: **9th Brazilian Conference on Intelligent Systems, BRACIS, Rio Grande do Sul, Brazil, October 20-23 (to appear)**. [S.l.: s.n.], 2020.
- TEAM, D. **Deepseek: Semantic Search for Large Scale Textual Data**. 2023. Available online: <<https://www.deepseek.ai/>>.
- TOUVRON, H. et al. **LLaMA: Open and Efficient Foundation Language Models**. 2023. Available at <<https://arxiv.org/abs/2302.13971>>.
- TURING, A. M. Computing machinery and intelligence. **Mind**, Oxford University Press, v. 59, n. 236, p. 433–460, 1950.
- UBERLÂNDIA, P. M. de. **Uberlândia conquista 1º lugar do país em serviços de cidades inteligentes**. 2021. <Https://www.uberlandia.mg.gov.br/2021/11/25/uberlandia-conquista-1o-lugar-do-pais-em-servicos-de-cidades-inteligentes>. Acesso em: 7 fev. 2024.
- _____. **Prefeitura de Uberlândia é a 1ª do país a homologar resultado de licitação no metaverso**. 2022. <Https://www.uberlandia.mg.gov.br/2022/09/02/prefeitura-de-uberlandia-e-a-1a-do-pais-a-homologar-resultado-de-licitacao-no-metaverso>. Acesso em: 7 fev. 2024.

_____. **Prefeitura de Uberlândia é a primeira do Brasil a realizar reunião de trabalho no metaverso.** 2022. [Https://www.uberlandia.mg.gov.br/2022/01/31/prefeitura-de-uberlandia-e-a-primeira-do-brasil-a-realizar-reuniao-de-trabalho-no-metaverso](https://www.uberlandia.mg.gov.br/2022/01/31/prefeitura-de-uberlandia-e-a-primeira-do-brasil-a-realizar-reuniao-de-trabalho-no-metaverso). Acesso em: 7 fev. 2024.

_____. **Uberlândia é a 1ª do Brasil no ranking Cidades com Serviços Inteligentes pelo segundo ano consecutivo.** 2022.

[Https://www.uberlandia.mg.gov.br/2022/12/16/uberlandia-e-a-1a-do-brasil-no-ranking-cidades-com-servicos-inteligentes-pelo-segundo-ano-consecutivo](https://www.uberlandia.mg.gov.br/2022/12/16/uberlandia-e-a-1a-do-brasil-no-ranking-cidades-com-servicos-inteligentes-pelo-segundo-ano-consecutivo). Acesso em: 7 fev. 2024.

VASWANI, A. et al. **Attention is All You Need.** 2017.

WANG, L. et al. Multilingual e5 text embeddings: A technical report. **arXiv preprint arXiv:2402.05672**, 2024.

XIAO, S. et al. **C-Pack: Packaged Resources To Advance General Chinese Embedding.** 2023.

ZHANG, T. et al. Bertscore: Evaluating text generation with bert. **International Conference on Learning Representations**, 2020.

ZHANG, X. et al. **Hierarchical Retrieval-Augmented Generation Model with Rethink for Multi-hop Question Answering.** 2024. Disponível em: <<https://arxiv.org/abs/2408.11875>>.

ZHU, X. et al. **Large Language Model Enhanced Text-to-SQL Generation: A Survey.** 2024. Disponível em: <<https://arxiv.org/abs/2410.06011>>.

Apêndices

APÊNDICE A

Interação do Chatbot com o Usuário

A seguir serão ilustradas algumas interações com usuários reais da Prefeitura Municipal de Uberlândia, onde o chat foi testado dentro de suas três funções básicas: Pesquisa em base de dados (planilha) a partir de nomes de candidatos, Pesquisa a partir de pergunta sobre o Estatuto dos servidores Públicos Municipais e Pesquisa em edital de Processo Seletivo para Contratação. Foi também inserida a ilustração da mensagem inicial do chatbot.

A partir da interação também é possível perceber o roteamento que é feito a partir da pergunta do usuário, endereçando a pergunta à base de dados correta.

The screenshot shows a WhatsApp-style chat interface. At the top left is a profile picture of a man named Abel. To the right are icons for a square, a downward arrow, and a magnifying glass. The message area has a light green background with a repeating pattern of school-related icons like books, pens, and globes.

Message 1:

Abel: 4.1.3):
- Ser doad...[Ler mais](#) 09:46 ✓

maria 10:08

Message 2:

Não encontrei exatamente 'Maria'. Encontrei candidatos similares:

- Maria Gonçalves
- Maria Julia Costela
- Maria Vitória Jesus
- Maria Eduarda das Neves
- Maria Cecília da Conceição

Por favor, seja mais específico com o nome completo. 10:08 ✓

maria gonçalves 10:09

Message 3:

Candidato encontrado: Maria Gonçalves

Informações:

- INSCRIÇÃO: 2767691
- NOME: Maria Gonçalves
- STATUS CONVOAÇÃO: Não Convocado
- EDITAL: Edital 10-2025

10:09 ✓

Requisitos para a vaga de professor de inglês 19:07

Conforme o Edital N° 01/2025 do Processo Seletivo Simplificado Público do Município de Uberlândia/MG, os **"requisitos para cargo de Professor de Inglês"** são baseados no contexto fornecido, especialmente no **"Anexo II - Conteúdo Programático"** (Contexto 3).

Requisitos para o Cargo de Professor de Inglês:

- *Formação Acadêmica*:**
 - **Ensino Superior**: O cargo exige formação de nível superior, conforme indicado na tabela de cargos do item 7.1 (Contexto 5), que lista os cargos de ensino superior, incluindo "Professor de Inglês".
- *Conhecimentos Avaliados***:
 - **Prova Objetiva**: Composta por:
 - **Conhecimentos Básicos (POCB)**:
 - Língua Portuguesa (10 questões, peso 2).
 - Legislação (10 questões, peso 2).
 - **Conhecimentos Específicos (POCE)**:
 - Conhecimentos específicos para o cargo de Professor de Inglês (15 questões, peso 4).
- *Conteúdo Programático Específico*** (Contexto 3 - Anexo II):
 - **Legislação**:
 - LDB (Lei nº 9.394/1996).
 - Constituição Federal (Capítulo III, Seção I - Educação).
 - ECA (Lei nº 8.069/1990).
 - Pareceres e resoluções do CNE/CEB (ex: Parecer nº 17/2001, Resolução nº 5/2009).
 - Lei Complementar Municipal nº 040/1992 (artigos 163 e 164).
 - Lei Municipal nº 9.626/2007 e alterações.
 - **Conhecimentos Específicos**:
 - Parte 1: Aborda psicologia do desenvolvimento, didática, metodologia de ensino, avaliação, inclusão, BNCC, entre outros.
 - Parte 2: Foco no ensino de língua inglesa, incluindo:
 - Didática e avaliação no ensino de língua estrangeira.
 - Interpretação de textos em inglês.
 - Aspectos gramaticais (tempos verbais, preposições, estruturas sintáticas).
 - BNCC aplicada ao ensino de inglês.
- *Sugestões Bibliográficas*** (Contexto 3):
 - Livros como "Modern English Grammar" (Alexander), "Cambridge Grammar of English" (Carter & McCarthy), e outros voltados para gramática e didática do inglês.
 - Documentos como a BNCC (seção 4.1.4.1 - Língua Inglesa no Ensino Fundamental).
- *Condições Gerais***:
 - O candidato deve atender aos requisitos do edital, como apresentar documentos válidos no dia da prova (item 8.6, Contexto 5) e seguir as regras de conduta durante o processo seletivo.

Observações:

- Não há menção a requisitos adicionais (como experiência profissional ou certificações específicas) no contexto fornecido.
- A prova é eliminatória e classificatória, com nota mínima de 60 pontos (item 9.3, Contexto 4).

Para mais detalhes, consulte o **"Anexo II"** do edital (Contexto 3) e a tabela de cargos (item 7.1, Contexto 5). 19:08 //

 Abel

Quais as vagas 19:06

Conforme o Edital Nº 01/2025 (alterado pela Retificação nº 01), as vagas disponíveis no Processo Seletivo Simplificado Público da Prefeitura de Uberlândia/MG são para os seguintes cargos:

Cargos de Nível Médio:

1. *Agente de Combate às Endemias*

Cargos de Nível Superior:

1. *Inspetor Escolar*
2. *Professor de Inglês*
3. *Professor de Matemática*
4. *Profissional de Educação Física*

Observações:

- As vagas são distribuídas entre ***ampla concorrência, **pessoas com deficiência*** (conforme Capítulo 5 do Edital) e ***candidatos autodeclarados negros*** (20% das vagas, conforme Lei Municipal nº 12.300/2015, Capítulo 6).
- O número exato de vagas por cargo não foi especificado no contexto fornecido. Para detalhes quantitativos, consulte o Anexo I ou tabelas complementares do Edital.

Fonte: Itens 7.1 (tabela de cargos) e 6.1 (reserva para negros).

Caso precise de informações sobre distribuição de vagas ou requisitos específicos, posso buscar no contexto complementar, se disponível.

19:06 ✓

The screenshot shows a messaging interface with a user profile picture and the name 'Abel'. The user asks 'Quantos dias de férias eu tenho' at 19:12. The bot responds with a detailed explanation about municipal public servants' vacation rights, citing Article 134 of the Statute of Public Servants of Uberlândia. It also mentions that vacations can be used in up to 3 periods of no less than 10 days each, if requested by the employee. The message from the bot includes a timestamp '19:12' and a blue checkmark icon.

Quantos dias de férias eu tenho 19:12

O servidor público municipal tem direito a ***30 (trinta) dias de férias por ano, conforme o **Artigo 134*** do Estatuto dos Servidores Públicos da Prefeitura de Uberlândia (Lei Municipal nº 40/1992), com redação dada pela Lei Complementar nº 634/2017.

A citação exata do artigo é:

"Art. 134 O servidor terá direito ao gozo de 30 (trinta) dias de férias a cada 12 (doze) meses de efetivo exercício, de acordo com escala organizada pelo titular do órgão de lotação."

Além disso, as férias podem ser usufruídas em até ***3 períodos não inferiores a 10 dias cada***, desde que haja interesse da administração e requerimento do servidor (§1º do mesmo artigo).

19:12 ✓

Ontem

Oi 17:01

Olá! 🙌 Eu sou o ***Agente de Integração Digital*** do Município de Uberlândia! 🏛️

Minha especialidade é auxiliar candidatos e ingressantes aos cargos efetivos da Prefeitura Municipal de Uberlândia. Estou aqui para tornar sua jornada mais tranquila e informativa! ✨

Como posso ajudar você hoje?

Aqui estão minhas funcionalidades:

1. 🔎 ***Status da Convocação***: Posso pesquisar pelo nome do candidato e verificar o status atual da sua convocação no concurso. Basta me dizer seu nome completo!

2. 💬 ***Dúvidas sobre o Edital nº 1/2025***: Tem alguma pergunta sobre as regras, datas ou procedimentos do edital? Me pergunte que eu esclareço para você!

3. ⚖️ ***Estatuto dos Servidores***: Precisa entender melhor seus direitos e deveres? Posso tirar dúvidas sobre a Lei Complementar nº 40/1992 (o Estatuto dos Servidores Públicos do Município).

17:01 ✓

APÊNDICE B

Perguntas e respostas cedidas pela banca organizadora de Concursos e Processos Seletivos

Pergunta: Fiz a inscrição para o processo seletivo porém o boleto não está sendo gerado...

Resposta: Prezado(a) candidato(a), Agradecemos seu contato, solicitamos que se atente aos itens abaixo do edital de abertura MUNICÍPIO DE UBERLÂNDIA/MG PROCESSO SELETIVO SIMPLIFICADO PÚBLICO EDITAL Nº 01/2025 (alterado conforme a Re-tificação nº 01)

3. DAS INSCRIÇÕES 3.1. As inscrições serão realizadas via Internet, no endereço eletrônico <<https://rboconcursos.selecao.net.br/>>, iniciando-se no dia 3 de fevereiro de 2025 e encerrando-se no dia 17 de fevereiro de 2025, observado o horário oficial de Brasília/DF e os itens estabelecidos no Capítulo 2. Das Condições para Inscrição, deste Edital. 3.2. Após o preenchimento da ficha de solicitação de inscrição on-line, o candidato deverá efetuar o pagamento da taxa de Inscrição, exclusivamente nas agências bancárias ou site/Aplicativo da Caixa Econômica Federal, Casas Lotéricas, em postos de autoatendimento da CEF, de acordo com o valor definido na Tabela 1, do Capítulo 1 deste Edital. 3.3. Objetivando evitar ônus desnecessários, o candidato deverá recolher o valor da taxa de inscrição somente se atender a todos os requisitos exigidos para o cargo pretendido. 3.4. A inscrição do candidato implicará no completo conhecimento e a tácita aceitação das normas legais pertinentes e condições estabelecidas neste Edital e seus Anexos, e as condições previstas em Lei, sobre as quais não poderá alegar desconhecimento. 3.5. O candidato deverá efetuar o pagamento da taxa de inscrição através de guia de arrecadação, pagável exclusivamente nas agências bancárias ou site/aplicativo da Caixa Econômica Federal, Casas Lotéricas, em postos de autoatendimento da CEF, com vencimento para o dia 18 de fevereiro de 2025, primeiro dia útil subsequente após a data de encerramento do período de inscrição. Em caso de feriado ou evento que acarrete o fechamento de agências

bancárias da Caixa Econômica Federal na localidade em que se encontra o candidato, a guia deverá ser paga antecipadamente. 3.5.1. A guia de arrecadação estará disponível para impressão após o preenchimento da ficha de solicitação de inscrição on-line no endereço eletrônico <<https://rboconcursos.selecao.net.br/>>, até o dia 18 de fevereiro de 2025, conforme cronograma do certame - Anexo V. 3.5.2. A inscrição somente será confirmada após a comprovação do pagamento da taxa de inscrição. Atenciosamente, SAC RBO

Pergunta: Realizei o pagamento e até o momento não houve atualização na aba da minha matrícula para o concurso.

Resposta: Prezado(a) candidato(a), Agradecemos seu contato, solicitamos que se atente aos itens abaixo do edital de abertura MUNICÍPIO DE UBERLÂNDIA/MG PROCESSO SELETIVO SIMPLIFICADO PÚBLICO EDITAL Nº 01/2025 (alterado conforme a Retificação nº 01)

3.12. A partir do dia 25 de fevereiro de 2025, o candidato deverá conferir nos endereços eletrônicos <<https://rboconcursos.selecao.net.br/>> e <<http://www.uberlandia.mg.gov.br>> se os dados da inscrição, efetuada via Internet, e se o valor da inscrição foi recebido pela RBO Assessoria Pública e Projetos Municipais Ltda., ou seja, se a inscrição está confirmada. **3.12.1.** Em caso negativo, o candidato deverá entrar com recurso, em até 2 (dois) dias úteis da data de publicação, nos termos do Capítulo 11 deste Edital. Atenciosamente, SAC RBO

Pergunta: Paguei o boleto, não acho o local da prova.

Resposta: Prezado(a) candidato(a), Agradecemos seu contato, solicitamos que se atente aos itens abaixo do edital de abertura MUNICÍPIO DE UBERLÂNDIA/MG PROCESSO SELETIVO SIMPLIFICADO PÚBLICO EDITAL Nº 01/2025 (alterado conforme a Retificação nº 01)

8.1. As provas objetivas serão realizadas na cidade de Uberlândia/MG, na data prevista de 16 de março de 2025, de acordo com a divisão dos períodos estabelecidos no item 8.1.1 deste capítulo, em locais e horários a serem comunicados oportunamente por meio do Edital de Convocação para as Provas Objetivas a ser publicado no Diário Oficial do Município de Uberlândia e divulgado através da Internet nos endereços eletrônicos <<https://rboconcursos.selecao.net.br/>> e <<http://www.uberlandia.mg.gov.br>>. observado o horário oficial de Brasília/DF. Atenciosamente, SAC RBO

Pergunta: Como faço para solicitar a isenção?

Resposta: Prezado(a) candidato(a), Agradecemos seu contato, solicitamos que se atente aos itens abaixo do edital de abertura MUNICÍPIO DE UBERLÂNDIA/MG PROCESSO SELETIVO SIMPLIFICADO PÚBLICO EDITAL Nº 01/2025 (alterado conforme a Re-

tificação nº 01)

4.2. O candidato que preencher a condição estabelecida no item 4.1.1, 4.1.2 ou 4.1.3 deverá solicitar a isenção do pagamento do valor de inscrição obedecendo aos seguintes procedimentos: **4.2.1.** Imprimir o formulário de solicitação de isenção da taxa de inscrição, constante no Anexo IV deste Edital, preenchê-lo e assiná-lo. **4.2.2.** Acessar no período de 3 a 4 de fevereiro de 2025, o site da RBO Assessoria Pública e Projetos Municipais Ltda. - <<https://rboconcursos.selecao.net.br/>> e, na área do Processo seletivo nº 01/2025 do Município de Uberlândia, realizar a inscrição online, optando pela isenção da taxa; **4.2.3.** Enviar o formulário especificado no subitem 4.2.1, juntamente com os documentos descritos nas alíneas do subitem 4.1.1.1. Atenciosamente, SAC RBO

Pergunta: Como envio os documentos da isenção?

Resposta: Prezado(a) candidato(a), Agradecemos seu contato, solicitamos que se atente aos itens abaixo do edital de abertura MUNICÍPIO DE UBERLÂNDIA/MG PROCESSO SELETIVO SIMPLIFICADO PÚBLICO EDITAL Nº 01/2025 (alterado conforme a Retificação nº 01)

4.2.3. Enviar o formulário especificado no subitem 4.2.1, juntamente com os documentos descritos nas alíneas do subitem 4.1.1.1. 4.1.2.1 ou 4.1.3.1 (conforme o caso), por meio de upload em campo específico na ficha de inscrição. **4.2.3.1.** Há apenas um único campo para upload da documentação indicada no item anterior, portanto os documentos deverão ser digitalizados e compactados (zipados) em um único arquivo de extensão zip antes do upload, com tamanho máximo TOTAL de 5MB. **4.2.3.2.** Antes de realizar o upload do arquivo zipado (contendo os arquivos relacionados no item 4.2.3 deste capítulo), verifique-o atentamente, visto que não haverá possibilidade de substituição de arquivo já enviado ou novo envio. **4.3.** O formulário discriminado no subitem 4.2.1 deverá ser encaminhado devidamente preenchido e conter a assinatura do solicitante e a data. **4.4.** Deverá ser realizada uma inscrição online e um envio de formulário com documentação comprobatória de isenção da taxa para cada cargo de interesse. Atenciosamente, SAC RBO

Pergunta: Posso me inscrever para dois cargos?

Resposta: Prezado(a) candidato(a), Agradecemos seu contato, solicitamos que se atente aos itens abaixo do edital de abertura MUNICÍPIO DE UBERLÂNDIA/MG PROCESSO SELETIVO SIMPLIFICADO PÚBLICO EDITAL Nº 01/2025 (alterado conforme a Retificação nº 01)

3.6. O candidato poderá efetuar até 2 (duas) inscrições neste Processo Seletivo, sendo 1 (uma) para cada período de aplicação de provas disposto na tabela do item 8.1.1, Capítulo 8 deste Edital. **3.6.1.** Em caso de mais de uma inscrição para o mesmo período de aplicação de prova, o candidato deverá optar somente por um cargo por período, sendo

considerado como ausente para os demais cargos do referido período de aplicação, mesmo que a aplicação das provas ocorra na mesma sala. **3.6.2.** Ocorrendo a hipótese do item 3.6.1, não haverá restituição parcial ou integral dos valores pagos a título de taxa de inscrição. Atenciosamente, SAC RBO

Pergunta: Eu sou negro, como me inscrevo na reserva, como faço para enviar a Declaração de negro e pardo?

Resposta: Prezado(a) candidato(a), Agradecemos seu contato, solicitamos que se atente aos itens abaixo do edital de abertura MUNICÍPIO DE UBERLÂNDIA/MG PROCESSO SELETIVO SIMPLIFICADO PÚBLICO EDITAL Nº 01/2025 (alterado conforme a Retificação nº 01)

6.4. O candidato deverá na ficha de inscrição, informar qual a cor se autodeclara, preto ou pardo, caso tenha interesse em participar das vagas reservadas. **6.4.1.** A autodeclaração somente terá validade se efetuada no momento da inscrição online. Atenciosamente, SAC RBO

Pergunta: Sou deficiente, como encaminho o laudo médico?

Resposta: Prezado(a) candidato(a), Agradecemos seu contato, solicitamos que se atente aos itens abaixo do edital de abertura MUNICÍPIO DE UBERLÂNDIA/MG PROCESSO SELETIVO SIMPLIFICADO PÚBLICO EDITAL Nº 01/2025 (alterado conforme a Retificação nº 01)

5.9. O candidato que desejar concorrer às vagas reservadas às pessoas com deficiência deverá declarar a condição na Ficha de Inscrição e proceder de acordo com o item 5.10.

5.10. O candidato com deficiência deverá encaminhar, por meio de upload em campo específico na ficha de inscrição, os documentos a seguir: **a)** Declaração constante no Anexo III deste Edital, devidamente preenchida e assinada pelo candidato, especificando a condição especial para a realização da prova (caso a condição especial seja necessária); e **b)** Laudo Médico, atestando a espécie e o grau ou nível de deficiência, com expressa referência ao código correspondente da Classificação Internacional de Doença CID, bem como a provável causa da deficiência, inclusive para assegurar previsão de adaptação da sua prova, informando o seu nome, número do RG e do CPF. Atenciosamente, SAC RBO

Pergunta: Sou deficiente, preciso de ajuda para fazer a prova objetiva.

Resposta: Prezado(a) candidato(a), Agradecemos seu contato, solicitamos que se atente aos itens abaixo do edital de abertura MUNICÍPIO DE UBERLÂNDIA/MG PROCESSO SELETIVO SIMPLIFICADO PÚBLICO EDITAL Nº 01/2025 (alterado conforme a Retificação nº 01)

3.16. O candidato que necessitar de condições especiais para realização das provas

deverá encaminhar, por meio de upload em campo específico na ficha de inscrição, declaração constante no Anexo III deste Edital, devidamente preenchida e assinada pelo candidato, especificando a condição especial para a realização da prova. **3.16.1.** O candidato que não o fizer durante o período de inscrição estabelecido no item anterior, não terá a prova e as condições especiais providenciadas, seja qual for o motivo alegado. **3.16.2.** O atendimento às condições solicitadas ficará sujeito à análise de viabilidade e razoabilidade do pedido. Atenciosamente, SAC RBO

Pergunta: Como fico sabendo se passei no concurso?

Resposta: Prezado(a) candidato(a), Agradecemos seu contato, solicitamos que se atente aos itens abaixo do edital de abertura MUNICÍPIO DE UBERLÂNDIA/MG PROCESSO SELETIVO SIMPLIFICADO PÚBLICO EDITAL Nº 01/2025 (alterado conforme a Retificação nº 01)

10.1. A nota final de cada candidato será IGUAL ao total de pontos obtidos na prova objetiva. **10.2.** Os candidatos serão classificados por ordem decrescente, da Nota Final, em lista de classificação por opção de cargo. **10.3.** Serão elaboradas 3 (três) listas de classificação, uma geral com a relação de todos os candidatos, incluindo os candidatos com deficiência e autodeclarados negros, uma específica, com a relação apenas dos candidatos com deficiência e outra específica com a relação apenas dos candidatos autodeclarados negros. **10.4.** O resultado preliminar do Processo Seletivo será publicado via Comunicado no Diário Oficial do Município de Uberlândia e divulgado na Internet nos endereços eletrônicos <<https://rboconcursos.selecao.net.br/>> e <<http://www.uberlandia.mg.gov.br>>. cabendo recurso nos termos do Capítulo 11. Dos Recursos deste Edital. **10.5.** Após o julgamento dos recursos, eventualmente interpostos, será divulgada a lista de Classificação Final, não cabendo mais recursos. **10.6.** A lista de Classificação Final será divulgada nos endereços eletrônicos <<https://rboconcursos.selecao.net.br/>> e <<http://www.uberlandia.mg.gov.br>>. **10.6.1.** Serão publicados no Diário Oficial do Município de Uberlândia apenas os resultados dos candidatos que lograram classificação no Processo Seletivo. **10.6.2.** O resultado geral final do Processo Seletivo poderá ser consultado nos endereços eletrônicos <<https://rboconcursos.selecao.net.br/>> e <<http://www.uberlandia.mg.gov.br>>. a contar da data de sua publicação no Diário Oficial do Município de Uberlândia. Atenciosamente, SAC RBO

Pergunta: Fiz a prova, mas minha nota está errada.

Resposta: Prezado(a) candidato(a), Agradecemos seu contato, solicitamos que se atente aos itens abaixo do edital de abertura MUNICÍPIO DE UBERLÂNDIA/MG PROCESSO SELETIVO SIMPLIFICADO PÚBLICO EDITAL Nº 01/2025 (alterado conforme a Retificação nº 01)

11.1.1. Todos os recursos deverão ser interpostos em até 2 (dois) dias úteis, a contar

da divulgação, por edital, de cada evento. **11.1.2.** Os recursos interpostos que não se refiram especificamente aos eventos aprazados não serão apreciados. **11.2.** Para recorrer, o candidato deverá utilizar o endereço eletrônico da empresa realizadora do certame <<https://rboconcursos.selecao.net.br/>>, acessar a área do Município de Uberlândia Edital Nº 01/2025 (Processo Seletivo) e seguir as instruções ali contidas. **11.3.** Somente serão apreciados os recursos interpostos dentro do prazo pré-estabelecido e que possuírem fundamentação e argumentação lógica e consistente que permita sua adequada avaliação. **11.4.** Não serão aceitos recursos interpostos por fax, e-mail, telegrama ou outro meio que não seja o especificado no item 11.2. **11.5.** Recurso inconsistente ou intempestivo será preliminarmente indeferido. **11.6.** Admitir-se-á um único recurso por candidato, para cada evento. Em caso de recurso referente ao gabarito, o candidato poderá interpor um recurso por questão. **11.6.1.** O recurso deverá ser individual, devidamente fundamentado e conter a identificação do Processo, nome do candidato, número de inscrição, cargo, o questionamento e o número da questão (se for o caso). **11.7.** Recebido o recurso, a RBO decidirá pelo provimento ou não do ato recorrido, dando-se ciência da referida decisão ao interessado por meio do endereço eletrônico <<https://rboconcursos.selecao.net.br/>>, na área do Município de Uberlândia - Edital Nº 01/2025 (Processo Seletivo). **11.8.** Em nenhuma hipótese serão aceitos pedidos de revisão de recurso, recursos de recursos e recurso de gabarito oficial definitivo. Atenciosamente, SAC RBO

Pergunta: Daqui a 2 dias será a prova, eu paguei o boleto, meu nome não consta na lista divulgada. O que irão fazer?

Resposta: Prezado(a) candidato(a), Agradecemos seu contato, solicitamos que se atente aos itens abaixo do edital de abertura MUNICÍPIO DE UBERLÂNDIA/MG PROCESSO SELETIVO SIMPLIFICADO PÚBLICO EDITAL Nº 01/2025 (alterado conforme a Retificação nº 01)

8.8. No dia da realização das provas, na hipótese de o candidato não constar nas listagens oficiais relativas aos locais de prova estabelecidos no Edital de Convocação, a RBO Assessoria Pública e Projetos Municipais Ltda., procederá à inclusão do referido candidato, através de preenchimento de formulário específico mediante a apresentação do comprovante de inscrição. **8.8.1.** A inclusão de que trata o item 8.8, será realizada de forma condicional e será confirmada pela RBO Assessoria Pública e Projetos Municipais Ltda., na fase de Julgamento das Provas Objetivas, com o intuito de se verificar a pertinência da referida inclusão. **8.8.2.** Constatada a improcedência da inscrição de que trata o item 8.8, a mesma será automaticamente cancelada sem direito a reclamação, independentemente de qualquer formalidade, considerados nulos todos os atos dela decorrentes. Atenciosamente, SAC RBO

Pergunta: Fiz a inscrição no cargo errado. Quero o cancelamento e o reembolso do valor.

Resposta: Prezado(a) candidato(a), Agradecemos seu contato, solicitamos que se atente aos itens abaixo do edital de abertura MUNICÍPIO DE UBERLÂNDIA/MG PROCESSO SELETIVO SIMPLIFICADO PÚBLICO EDITAL Nº 01/2025 (alterado conforme a Retificação nº 01)

3.9. Salvo nos casos de suspensão, anulação ou cancelamento do certame, não haverá devolução, parcial ou integral, da importância paga, ainda que superior ou em duplicidade, nem isenção total ou parcial de pagamento do valor da taxa de inscrição, seja qual for o motivo alegado. **3.10.** Efetivada a inscrição, não serão aceitos pedidos de alteração de opção de cargo sob hipótese alguma, portanto, antes de efetuar o pagamento da taxa de inscrição, o candidato deve verificar atentamente a opção preenchida. Atenciosamente, SAC RBO

Pergunta: Quero trocar de cargo, fiz a inscrição no cargo errado.

Resposta: Prezado(a) candidato(a), Agradecemos seu contato, solicitamos que se atente aos itens abaixo do edital de abertura MUNICÍPIO DE UBERLÂNDIA/MG PROCESSO SELETIVO SIMPLIFICADO PÚBLICO EDITAL Nº 01/2025 (alterado conforme a Retificação nº 01)

3.10. Efetivada a inscrição, não serão aceitos pedidos de alteração de opção de cargo sob hipótese alguma, portanto, antes de efetuar o pagamento da taxa de inscrição, o candidato deve verificar atentamente a opção preenchida. Atenciosamente, SAC RBO

Pergunta: Eu posso fazer a prova em outra data, pois no dia 16 estarei fazendo outro concurso que já havia me inscrito antes?

Resposta: Prezado(a) candidato(a), Agradecemos seu contato, solicitamos que se atente aos itens abaixo do edital de abertura MUNICÍPIO DE UBERLÂNDIA/MG PROCESSO SELETIVO SIMPLIFICADO PÚBLICO EDITAL Nº 01/2025 (alterado conforme a Retificação nº 01)

8.2. Ao candidato só será permitida a participação nas provas, na respectiva data, horário e local a serem divulgados de acordo com as informações constantes no item 8.1 deste Capítulo. **8.3.** Não será permitida, em hipótese alguma, a realização das provas em outro dia, horário ou fora do local designado. Atenciosamente, SAC RBO

Pergunta: Eu passei no concurso, quero saber quando vão me chamar.

Resposta: Prezado(a) candidato(a), Agradecemos seu contato, solicitamos que se atente aos itens abaixo do edital de abertura MUNICÍPIO DE UBERLÂNDIA/MG PROCESSO SELETIVO SIMPLIFICADO PÚBLICO EDITAL Nº 01/2025 (alterado conforme a Retificação nº 01)

13.1. A aprovação e a classificação final do candidato fora do número de vagas ofer-
ta-

das neste Edital gera, para o candidato, apenas a expectativa de direito à contratação. O Município de Uberlândia reserva-se o direito de proceder às contratações em número que atenda ao interesse e às necessidades do serviço, dentro do prazo de validade do Processo Seletivo. **13.2.** A contratação dos candidatos que vierem a ser convocados pelo Município de Uberlândia será temporária, regida pelas normas do Direito Administrativo, pela Lei Municipal 9626 de 2007 e suas alterações, não se aplicando as normas contidas na Consolidação das Leis do Trabalho - CLT. **13.3.** É de responsabilidade e dever do candidato cumprir rigorosamente os prazos estabelecidos pelo Município de Uberlândia, sob pena de perda do direito da contratação. Atenciosamente, SAC RBO

Pergunta: Passei no concurso e preciso saber onde vou trabalhar.

Resposta: Prezado(a) candidato(a), Agradecemos seu contato, solicitamos que se atente aos itens abaixo do edital de abertura MUNICÍPIO DE UBERLÂNDIA/MG PROCESSO SELETIVO SIMPLIFICADO PÚBLICO EDITAL Nº 01/2025 (alterado conforme a Retificação nº 01)

1.2. As vagas oferecidas são para o município de Uberlândia/MG, porém os locais de trabalho serão definidos no ato da contratação e poderão ser alterados durante a vigência do contrato. Atenciosamente, SAC RBO

APÊNDICE C

Resultados do último experimento por modelo em cada métrica

A seguir os resultados de cada métrica utilizada no experimento feito com as perguntas cedidas pela banca realizadora, onde o modelo foi testado em 6 cenários.

