UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Desenvolvimento de um Sistema Baseado em RAG para Consulta e Extração de Informações em Documentos Técnicos

Filipe Augusto Valentins Araújo

Monografia - MBA em Inteligência Artificial e Big Data

ICMC USP SÃO CARLOS

Filipe Augusto Valentins Araújo

Desenvolvimento de um Sistema Baseado em RAG para Consulta e Extração de Informações em Documentos Técnicos

Monografia apresentada ao Departamento de Ciências de Computação do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo - ICMC/USP, como parte dos requisitos para obtenção do título de Especialista em Inteligência Artificial e Big Data.

Área de concentração: Inteligência Artificial

Orientador: Prof. Dr. Gesiel Rios Lopes

Versão original

São Carlos

2025

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi, ICMC/USP, com os dados fornecidos pelo(a) autor(a)

S856m	Filipe Augusto Valentins Araújo Desenvolvimento de um Sistema Baseado em RAG para Consulta e Extração de Informações em Documentos Técnicos / Filipe Augusto Valentins Araújo ; orientador Gesiel Rios Lopes. – São Carlos, 2025. 54 p. : il. (algumas color.) ; 30 cm. Monografia (MBA em Inteligência Artificial e Big Data) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2025. 1. LaTeX. 2. abnTeX. 3. Classe USPSC. 4. Editoração de texto. 5. Normalização da documentação. 6. Tese. 7. Dissertação. 8. Documentos (elaboração). 9. Documentos eletrônicos. I. Lopes, Gesiel Rios, orient. II. Título.

Filipe Augusto Valentins Araújo

Development of a RAG-Based System for Querying and Extracting Information from Technical Documents

Monograph presented to the Departamento de Ciências de Computação do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo - ICMC/USP, as part of the requirements for obtaining the title of Specialist in Artificial Intelligence and Big Data.

Concentration area: Artificial Intelligence

Advisor: Prof. Dr. Gesiel Rios Lopes

Original version

São Carlos

2025

AGRADECIMENTOS

Primeiramente, expresso minha sincera gratidão a toda a equipe da USP pelo constante suporte e pelo acolhimento durante esta etapa acadêmica. Aos meus pais, agradeço profundamente pelo apoio incondicional, incentivo e compreensão em todos os momentos desta jornada.

Estendo meus agradecimentos aos colegas da USP, que sempre se mostraram solícitos e colaborativos, contribuindo de maneira significativa para o desenvolvimento deste trabalho.

Agradeço também às ferramentas de Inteligência Artificial que me auxiliaram no processo de elaboração deste trabalho. O ChatGPT foi utilizado como apoio para a correção de textos e no desenvolvimento de trechos de código, enquanto o Gemini foi empregado na geração de imagens ilustrativas. Ressalto que tais ferramentas tiveram caráter exclusivamente de suporte, não substituindo a análise crítica nem a responsabilidade autoral deste trabalho.

Por fim, registro minha especial gratidão ao meu orientador, Prof. Gesiel, cuja orientação, dedicação e confiança foram fundamentais desde o início até a conclusão deste projeto.

"Os cientistas estudam o mundo como ele é,
os engenheiros criam um mundo como ele nunca havia sido"
Theodore von Karman

RESUMO

ARAÚJO, F. **Desenvolvimento de um Sistema Baseado em RAG para Consulta e Extração de Informações em Documentos Técnicos**. 2025. 54 p. Monografia (MBA em Inteligência Artificial e Big Data) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2025.

A adoção de Modelos de Linguagem de Grande Escala (LLMs) tem avançado rapidamente no setor industrial, mas sua aplicação em ambientes corporativos enfrenta barreiras relacionadas à segurança e ao risco de vazamento de informações sensíveis. Diante desse contexto, este trabalho propõe o uso da técnica de Retrieval-Augmented Generation (RAG) como estratégia para viabilizar a criação de modelos internos, capazes de operar de forma isolada em servidores locais, garantindo maior controle sobre os dados. O projeto contemplou a implementação de uma arquitetura RAG aplicada a documentos técnicos, além da comparação entre diferentes modelos de linguagem (Openai/gpt-oss-20b, openchat-3.5-0106, meta-llama-3.1-8b e Gemini 2.5 Flash), a fim de analisar sua aplicabilidade em cenários de uso corporativo. Como resultado, foi demonstrada a possibilidade de estruturar modelos locais sem dependência de serviços externos, o que abre caminho para futuras soluções de IA em ambientes empresariais restritos. Observou-se ainda a diferença significativa entre modelos de 8B e 20B parâmetros, evidenciando a necessidade de infraestrutura computacional robusta para sustentar agentes de qualidade em aplicações reais.

Palavras-chave:Recuperação da Informação. Modelos de Linguagem. Inteligência Artificial. RAG.

ABSTRACT

ARAÚJO, F. **Development of a RAG-Based System for Querying and Extracting Information from Technical Documents**. 2025. 54 p. Monograph (MBA in Artificial Intelligence and Big Data) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2025.

The adoption of Large Language Models (LLMs) has grown rapidly in industrial environments, but their use within corporate settings faces challenges related to security and the risk of sensitive information leakage. In this context, this work explores Retrieval-Augmented Generation (RAG) as a strategy to enable the development of internal models running entirely on local servers, ensuring greater data control. The project involved implementing a RAG-based architecture applied to technical documents, along with a comparative evaluation of different language models (Openai/gpt-oss-20b, openchat-3.5-0106, meta-llama-3.1-8b, and Gemini 2.5 Flash) to assess their applicability in corporate scenarios. The results demonstrated the feasibility of deploying local models without reliance on external services, paving the way for future AI solutions in restricted business environments. Furthermore, a clear performance gap was observed between 8B and 20B parameter models, highlighting the necessity of robust computational infrastructure to sustain high-quality agents for practical applications.

Keywords: Information Retrieval. Language Models. Artificial Intelligence. RAG.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE QUADROS

LISTA DE ABREVIATURAS E SIGLAS

AI	Inteligência Artificial (Artificial Intelligence)
GPT	Generative Pre-trained Transformer
IEC	International Electrotechnical Commission
IT	Instrução Técnica
LLM	Modelo de Linguagem de Grande Porte (Large Language Model)
NLP	Processamento de Linguagem Natural (Natural Language Processing)
PDF	Portable Document Format
RAG	Recuperação Aumentada por Geração (Retrieval-Augmented Generation)
TD	Transformer Design Standard
USP	Universidade de São Paulo
USPSC	Campus USP de São Carlos

SUMÁRIO

1 INTRODUÇÃO

Nos últimos anos, o uso de modelos de linguagem de grande porte (LLMs) tem se destacado como uma das abordagens mais eficazes para automatizar tarefas complexas relacionadas ao acesso e à interpretação de informações técnicas. (Naveed *et al.*, 2024) destacam que esses modelos são capazes de realizar tarefas que vão desde a geração de texto até a compreensão contextual de comandos em linguagem natural, o que os torna altamente versáteis em domínios especializados. Já (Yu; Zhang; Chu, 2025), ao analisarem o avanço dos modelos multimodais, reforçam que a estrutura dos LLMs permite não apenas lidar com linguagem escrita, mas também integrar outras modalidades informacionais, ampliando ainda mais seu potencial de aplicação em cenários industriais complexos.

A proposta deste trabalho ganha relevância dentro de um cenário industrial cada vez mais pressionado por demandas emergentes, como é o caso da crescente necessidade de infraestrutura energética para suportar o funcionamento de centros de dados responsáveis por manter aplicações baseadas em inteligência artificial. (Arslan *et al.*, 2024) ressaltam que a expansão desses centros demanda soluções de engenharia que sejam ao mesmo tempo robustas e ágeis, principalmente no que diz respeito ao fornecimento elétrico. De forma complementar, (Bora; Cuayáhuitl, 2024) evidenciam que o crescimento de aplicações computacionais de alto consumo intensifica a procura por transformadores de grande porte, consolidando a urgência por melhorias nos processos que envolvem seu projeto e fabricação.

Nesse contexto, ganha destaque a necessidade por ferramentas que otimizem os processos de desenvolvimento e projeto técnico, sobretudo nas etapas iniciais, como o acesso a normas, instruções e regulamentos específicos. (Paranhos; Souza; Almeida, 2024), ao avaliarem diferentes arquiteturas de RAG no domínio jurídico, demonstraram que a complexidade e extensão dos documentos técnicos dificultam o acesso eficiente à informação por parte dos usuários. De maneira semelhante, (Siriwardhana *et al.*, 2023) mostraram que, ao adaptar modelos baseados em RAG a domínios específicos, como medicina e saúde pública, é possível aumentar significativamente a precisão e a utilidade prática das respostas fornecidas por esses sistemas.

Além das barreiras estruturais dos documentos técnicos, outro desafio recorrente está ligado ao perfil dos profissionais que atuam nesses ambientes. Embora a indústria conte com um número crescente de profissionais em formação, muitos ainda enfrentam dificuldades ao lidar com vocabulário técnico e com a interpretação adequada de normas especializadas. (Naveed *et al.*, 2024) alertam que essa lacuna de qualificação pode comprometer a qualidade e a segurança de projetos complexos, tornando essencial a adoção de ferramentas que promovam maior clareza e rastreabilidade. Corroborando essa visão, (Bora; Cuayáhuitl,

2024) ressaltam que, ao integrar a técnica de RAG com bases especializadas, é possível mitigar erros de interpretação e acelerar o processo de acesso à informação técnica.

Nesse cenário, alguns trabalhos vêm se destacando por apresentar soluções alinhadas com essa problemática. (Kuratomi; Kawakubo; Falcão, 2025) propuseram um assistente virtual baseado em RAG para facilitar a navegação em documentos normativos da Universidade de São Paulo, oferecendo suporte contextualizado e com alta precisão em respostas voltadas à burocracia institucional. De forma semelhante, (Cho; Park; Um, 2024) desenvolveram um sistema especializado em manuais de operação de máquinas-ferramenta, voltado para ambientes industriais, com o intuito de reduzir erros operacionais e aumentar a acessibilidade ao conhecimento técnico, mesmo por profissionais com menor nível de especialização.

É nesse ponto que a presente proposta se insere, ao investigar e propor uma solução baseada em LLMs e RAG para apoiar projetistas e engenheiros na obtenção rápida, qualificada e precisa de informações técnicas voltadas ao dimensionamento e construção de transformadores de grande porte. Conforme observado por (Siriwardhana *et al.*, 2023), a adaptação de modelos RAG a domínios específicos permite não apenas reduzir o tempo de resposta, mas também garantir consistência e fidelidade às fontes normativas. (Arslan *et al.*, 2024), por sua vez, defendem que esse tipo de arquitetura oferece uma base sólida para aplicações práticas em ambientes técnicos e industriais, como o proposto neste TCC.

Além das abordagens mencionadas, algumas soluções acadêmicas recentes reforçam a aplicabilidade de modelos RAG e LLMs em domínios altamente técnicos. Silva (Silva, 2025) propôs o uso de um assistente virtual para coordenadores de curso, aplicando técnicas de RAG para a gestão de perguntas frequentes, com resultados positivos quanto à contextualização e à robustez das respostas. Essa experiência demonstra o potencial da tecnologia mesmo em contextos com vocabulário técnico específico e alta necessidade de precisão, características igualmente presentes nos ambientes industriais abordados neste TCC. Da mesma forma, Garcia, Silva e Oliveira (Garcia; Silva; Oliveira, 2025) desenvolveram um chatbot acadêmico utilizando modelos de linguagem, com foco em recuperar normas e informações institucionais — um cenário muito semelhante à recuperação de instruções técnicas na engenharia.

Deus García (García, 2025) também contribui com a discussão ao relatar a criação de um sistema voltado à recuperação de documentação empresarial, integrando IA generativa e técnicas RAG. Seu projeto enfatiza a eficiência na consulta em linguagem natural e a importância da integração com sistemas já existentes, como o Microsoft Teams, o que ressalta a relevância da acessibilidade e adoção das ferramentas no ambiente de trabalho. Complementarmente, o trabalho de Jaguán (Jaguán, 2025) aprofunda a aplicação de modelos RAG em sistemas multiagente voltados ao tráfego aéreo, um setor igualmente crítico e sensível a falhas de interpretação técnica. Ambas as abordagens reforçam a

viabilidade de soluções baseadas em LLMs em setores onde a confiabilidade da informação é essencial — sustentando a proposta deste TCC de aplicar essa mesma lógica ao setor de engenharia elétrica de alta complexidade.

1.1 Objetivos

1.1.1 *Objetivo geral*

Desenvolver uma solução computacional baseada em modelos de linguagem de grande porte (LLMs), integrados à técnica de Recuperação Aumentada por Geração (RAG), com o propósito de auxiliar projetistas na extração ágil, precisa e rastreável de informações técnicas específicas contidas em documentos normativos, como Instruções Técnicas (ITs) e Termos de Desenho (TDs), no contexto da construção de transformadores de grande porte. A proposta contempla a estruturação de uma base documental técnica, a formulação de consultas em linguagem natural, a seleção e avaliação de diferentes modelos de IA e a validação do sistema por profissionais da área de projetos, assegurando a viabilidade de sua utilização em ambientes internos e protegidos contra vazamento de informações corporativas.

1.1.2 *Objetivos específicos*

I Identificar os principais desafios enfrentados por projetistas na busca por informações em Instruções Técnicas (ITs) e normas aplicáveis ao setor de transformadores;

II Estruturar uma base de dados contendo documentos técnicos relevantes à rotina de projeto de transformadores de grande porte;

III Implementar um sistema baseado na arquitetura RAG, capaz de processar consultas em linguagem natural e retornar trechos específicos de documentos técnicos;

IV Elaborar um conjunto de perguntas que aborde os temas presentes nos documentos técnicos selecionados;

V Selecionar diferentes modelos de inteligência artificial para responder às perguntas elaboradas;

VI Definir perfis distintos de profissionais da área de projetos, com capacidade técnica para validar as respostas fornecidas pelos modelos LLM;

VII Avaliar a qualidade e a confiabilidade das respostas geradas pelo sistema, considerando critérios como precisão, relevância e rastreabilidade das informações;

VIII Verificar a viabilidade de utilização da ferramenta em ambiente interno (on-premises), assegurando que os dados corporativos não sejam expostos à rede pública.

1.2 Justificativa

A justificativa deste projeto fundamenta-se na necessidade concreta de modernizar e agilizar o acesso a informações técnicas em ambientes industriais de alta complexidade. Transformadores de grande porte são componentes críticos na infraestrutura energética mundial, e qualquer falha no seu processo de projeto ou construção pode acarretar consequências significativas, tanto em termos econômicos quanto operacionais. Portanto, garantir que os profissionais envolvidos tenham acesso rápido e confiável a normas e instruções específicas é um fator determinante para a qualidade e a segurança do produto final.

Adicionalmente, a elevada rotatividade de profissionais e a carência de mão de obra especializada tornam ainda mais evidente a lacuna entre o conhecimento técnico disponível nos documentos normativos e a capacidade prática de interpretá-los corretamente no dia a dia. Ferramentas que promovam o entendimento e a aplicação segura dessas normas, especialmente se baseadas em abordagens modernas como a Recuperação Aumentada por Geração (RAG), têm potencial para transformar a rotina de trabalho e promover ganhos reais de produtividade.

A motivação pessoal que impulsiona este projeto está relacionada à vivência prática no ambiente de engenharia e à observação direta das dificuldades enfrentadas por colegas de trabalho e projetistas experientes na busca por respostas objetivas em meio a documentos técnicos extensos. Muitas vezes, o tempo que poderia ser direcionado à etapa criativa e decisória do projeto é consumido em tarefas repetitivas de leitura e interpretação normativa. Assim, surge a motivação para aplicar conhecimentos em inteligência artificial de forma prática e orientada à resolução de problemas reais da indústria.

2 FUNDAMENTAÇÃO TEÓRICA

2.1 Large Language Models (LLMs)

Os Large Language Models (LLMs) são modelos de inteligência artificial desenvolvidos com o objetivo de compreender e gerar linguagem natural de forma contextualizada e coerente. A base desses modelos está no uso da arquitetura Transformer, apresentada por Vaswani et al. (2017), que revolucionou o campo do Processamento de Linguagem Natural (NLP) ao permitir o tratamento de longas sequências textuais com alta eficiência. Os LLMs são treinados com grandes volumes de dados textuais retirados de livros, artigos, sites e documentos técnicos, permitindo que aprendam padrões linguísticos, sintaxe, semântica e até mesmo especializações de determinados domínios. Modelos como GPT (Generative Pre-trained Transformer), BERT, LLaMA e Mistral são exemplos de LLMs que podem ser adaptados para diversas finalidades, como geração de texto, tradução, sumarização e resposta a perguntas. No contexto técnico-industrial, a aplicação de LLMs se torna relevante ao permitir que profissionais possam interagir com bases complexas de informação por meio de perguntas em linguagem natural, obtendo como retorno explicações claras e direcionadas sobre conteúdos normativos ou procedimentos técnicos.

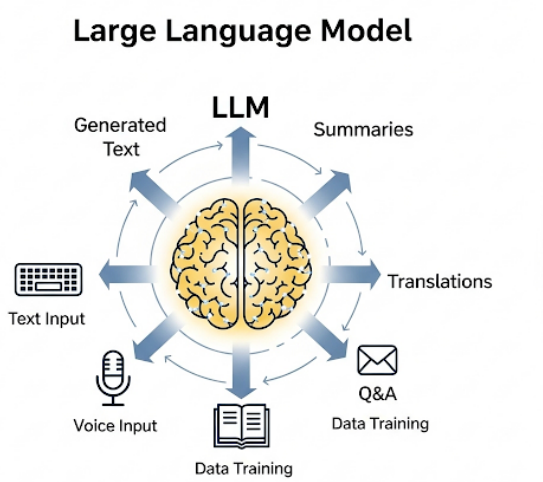

Figura 1 – Representação ilustrativa de um Modelo de Linguagem de Grande Escala (LLM).

Fonte: Elaborado pelo autor, 2025.

A Figura 1 ilustra, de forma esquemática, o funcionamento conceitual de um LLM, evidenciando sua arquitetura em camadas e a relação entre dados de entrada e geração de respostas.

2.2 Retrieval-Augmented Generation (RAG)

O método Retrieval-Augmented Generation (RAG) surgiu como uma resposta às limitações dos modelos de linguagem convencionais que, apesar de poderosos, ainda enfrentam dificuldades em oferecer respostas com base em informações atualizadas, específicas ou de fontes confiáveis. O RAG combina dois componentes: um módulo de recuperação de informação (retriever), responsável por buscar documentos relevantes em uma base previamente indexada, e um módulo de geração de texto (generator), que utiliza essas informações recuperadas para gerar uma resposta textual fundamentada. Essa arquitetura se destaca por permitir que os LLMs "consultem" diretamente um repositório de conhecimento externo, como uma coleção de documentos técnicos em PDF, repositórios de normas, artigos científicos ou manuais. A geração textual, nesse modelo, é baseada nos trechos reais recuperados, o que garante rastreabilidade da resposta, maior precisão técnica e reduz drasticamente os riscos de "alucinação" — quando o modelo fornece informações incorretas ou inventadas. Por essa razão, o RAG tem sido amplamente adotado em domínios que exigem alto grau de confiabilidade, como medicina, direito e engenharia.

2.3 Instruções Técnicas (ITs)

As Instruções Técnicas (ITs) são documentos elaborados com o objetivo de padronizar processos, estabelecer critérios técnicos, registrar metodologias e garantir conformidade com normas regulatórias, especialmente em setores industriais de alta complexidade. No contexto da engenharia elétrica, e mais especificamente no projeto e fabricação de transformadores de grande porte, as ITs orientam desde o cálculo e o dimensionamento das bobinas, núcleos e isolamentos, até os testes finais de conformidade com normas como a ABNT NBR ou IEC. Apesar de sua importância, as ITs tendem a ser documentos extensos, densos e redigidos com linguagem altamente técnica. Esse fator, aliado à constante atualização dos procedimentos e da legislação, dificulta o acesso rápido às informações pontuais por parte dos projetistas e engenheiros. Assim, a integração entre as ITs e sistemas baseados em RAG e LLMs representa um avanço significativo, ao permitir que o conhecimento técnico contido nesses documentos seja acessado de forma ágil, interativa e com base na demanda real de cada consulta.

2.4 Estruturação de um Projeto Baseado em RAG para Consultas Técnicas

A criação de um sistema computacional com base no método RAG para consulta de informações técnicas em ITs envolve uma série de etapas estruturadas, que abrangem desde o preparo dos dados até a aplicação prática em ambiente industrial. A seguir, descrevem-se os principais passos:

2.4.1 *Coleta e Organização dos Documentos Técnicos*

O primeiro passo é identificar e reunir os documentos normativos e técnicos que farão parte do repositório de consulta. No caso deste TCC, são utilizados documentos que abordam assuntos distintos como normas de segurança, cálculos de tração e modulo elástico (ou young) com o intuito de simular as Instruções Técnicas (ITs) da área de engenharia de transformadores. Esses documentos podem estar em formatos variados (PDF, DOCX, etc.), sendo necessário convertê-los para texto estruturado (plain text ou JSON) para fins de processamento posterior.

2.4.2 *Segmentação e Indexação dos Dados*

Com os textos extraídos, aplica-se um processo de chunking, onde os documentos são divididos em pequenos trechos (parágrafos, seções ou frases), facilitando a indexação semântica. Cada trecho recebe um identificador único e pode ser armazenado em um banco vetorial (vector store), onde são criadas representações matemáticas dos conteúdos por meio de embeddings gerados por modelos como Sentence-BERT, FAISS ou ChromaDB.

2.4.3 *Configuração do Módulo de Recuperação (Retriever)*

Nesta etapa, implementa-se o mecanismo de recuperação semântica que, ao receber uma pergunta do usuário, é capaz de localizar os trechos mais relevantes na base de documentos indexados. A similaridade entre a pergunta e os trechos é calculada por meio de medidas vetoriais (como cosseno ou dot product).

2.4.4 *Integração com o Modelo de Linguagem (LLM)*

O modelo de linguagem é responsável por ler a pergunta do usuário e, com base nos documentos recuperados, gerar uma resposta coerente e técnica. É nesta etapa que ocorre a geração aumentada, pois o modelo não se baseia unicamente no que foi treinado, mas nas informações reais extraídas das documentações.

2.4.5 *Seleção e utilização dos modelos de LLM*

Existem diferentes formas de utilizar um modelo de linguagem de grande porte (LLM). Atualmente, é possível acessá-los por meio de APIs de provedores externos ou através do download de versões locais com diferentes quantidades de parâmetros — como modelos de 8 bilhões (8B) ou 20 bilhões (20B). O número de parâmetros está diretamente relacionado à capacidade de armazenamento de contexto e à sofisticação das respostas geradas. No entanto, modelos com maior quantidade de parâmetros exigem infraestrutura computacional mais robusta para sua execução local, embora ofereçam, em contrapartida, ganhos significativos em precisão, coerência e performance.

2.4.6 *Validação e Testes*

Por fim, são realizados testes de validação do sistema com perguntas formuladas com base nos conteúdos presentes nos documentos utilizados, os quais simulam situações recorrentes na rotina do setor de projetos. Avalia-se a precisão das respostas, a clareza da linguagem gerada e a aderência às normas técnicas. A etapa de avaliação é conduzida por profissionais com experiência na área, responsáveis por mensurar a qualidade, confiabilidade e adequação das respostas fornecidas por cada modelo LLM.

3 METODOLOGIA

Para o desenvolvimento deste trabalho, foi implementado um sistema computacional baseado na arquitetura RAG (Retrieval-Augmented Generation), com o intuito de auxiliar profissionais da engenharia na consulta rápida e precisa a documentos técnicos, como Instruções Técnicas (ITs). A metodologia adotada compreende quatro etapas principais: a preparação e indexação dos documentos técnicos; a configuração do módulo de recuperação semântica integrado a diferentes modelos de linguagem de grande porte (LLMs); a elaboração de um conjunto de perguntas relacionadas aos conteúdos dos documentos; e, por fim, a validação das respostas geradas por cada modelo, com base em critérios de precisão, clareza e aderência às normas técnicas.

3.1 Preparação do Ambiente de Desenvolvimento

A etapa inicial consistiu na configuração do ambiente de desenvolvimento, realizada por meio da instalação do *Anaconda*, ferramenta que possibilita a criação e o gerenciamento de ambientes virtuais. Dentro desse ambiente dedicado, foram instaladas as bibliotecas necessárias ao projeto, incluindo aquelas não nativas do *Python*, mas indispensáveis para o processamento e manipulação dos documentos técnicos. Com a configuração concluída, tornou-se possível inicializar o *Jupyter Notebook* com todas as dependências previamente preparadas, garantindo maior organização e reprodutibilidade ao experimento.

Paralelamente, utilizou-se o software *LM Studio*, responsável pela instalação e execução de Modelos de Linguagem de Grande Porte (LLMs) de forma local e gratuita. Além de permitir a utilização direta desses modelos, o *LM Studio* oferece a funcionalidade de criação de *endpoints* em servidor local, possibilitando que requisições sejam feitas a partir do código implementado no *Jupyter Notebook*. Dessa forma, tornou-se viável a integração de diferentes agentes de linguagem, como o *Ollama*, ampliando as possibilidades de experimentação dentro do escopo do projeto.

3.2 Coleta e Processamento de Documentos

Foram selecionados documentos técnicos em formato PDF, simulando Instruções Técnicas (ITs) com conteúdos voltados para três temas específicos: *NBR35*, Tração e Módulo de Young. Esses documentos serviram como base para compor o repositório de informações a ser consultado pelo sistema proposto.

Os documentos, majoritariamente em formato PDF, foram processados por um módulo desenvolvido em Python, utilizando a biblioteca *PyMuPDF* para extração textual. Após a extração, o conteúdo foi segmentado em trechos sobrepostos (chunks) com auxílio

da ferramenta *RecursiveCharacterTextSplitter*, facilitando a indexação semântica e posterior recuperação das informações. Os textos segmentados foram armazenados com seus respectivos metadados, compondo uma base estruturada de conhecimento para consulta automatizada.

3.3 Indexação Vetorial e Recuperação Semântica

Com os documentos segmentados, aplicou-se um modelo de embeddings para transformar os chunks textuais em vetores numéricos. Para essa tarefa, foi utilizado o modelo *all-MiniLM-L6-v2* da biblioteca *HuggingFace*, por apresentar bom equilíbrio entre desempenho e custo computacional. Os vetores gerados foram indexados em uma base vetorial utilizando o algoritmo *FAISS*, possibilitando a recuperação dos trechos mais semanticamente próximos à consulta do usuário. A similaridade entre pergunta e contexto foi calculada com base na distância de cosseno entre os vetores.

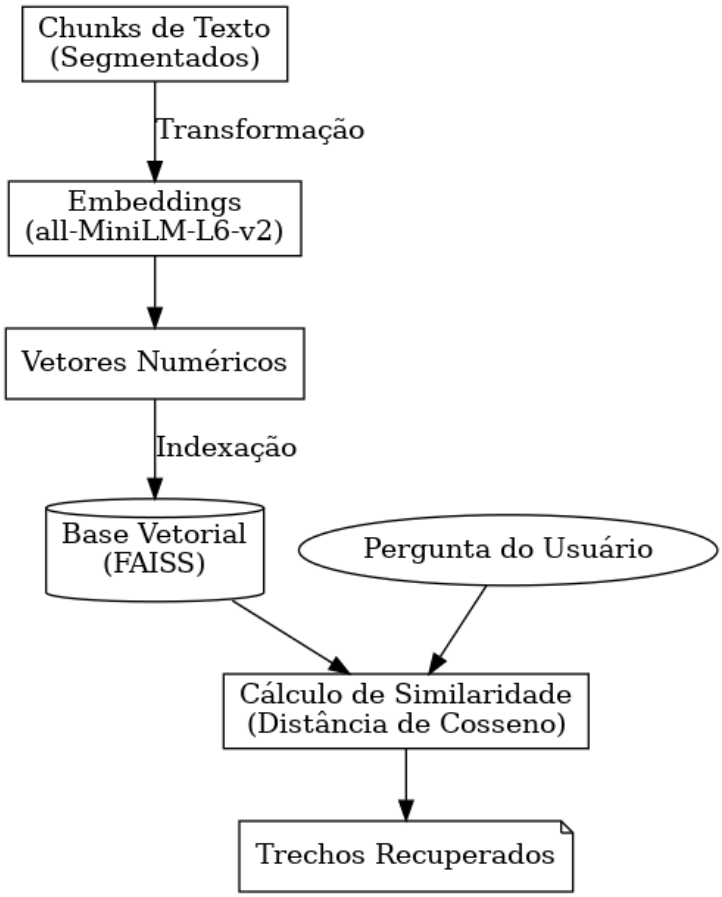

Figura 2 – Fluxograma da indexação vetorial e recuperação semântica.

Fonte: Elaborado pelo autor, 2025.

3.4 Integração com Modelos de Linguagem)

O módulo de recuperação foi integrado a diferentes Modelos de Linguagem de Grande Porte (LLMs). Cada consulta, formulada em linguagem natural, foi enriquecida

com os trechos recuperados da base documental e submetida ao modelo selecionado, o qual, a partir desse contexto adicional, gerava uma resposta mais precisa e fundamentada. Esse fluxo caracteriza a integração plena entre as etapas de *retrieval* e *generation*, que compõem a essência da arquitetura RAG.

Para a validação da proposta, foram empregados quatro modelos distintos: *Gemini 1.5 Flash*, *LLaMA 3.1-8B*, *OpenChat 3.5* e *GPT-OSS-20B*. O *Gemini 1.5 Flash* foi acessado por meio de API, utilizando chave fornecida gratuitamente a partir de uma conta Google. Já os demais modelos foram executados localmente, com auxílio do *LM Studio*, que permitiu a instalação, gerenciamento e disponibilização dos agentes por meio de *endpoints* em servidor local.

A quantidade de parâmetros suportada por cada LLM, bem como suas respectivas organizações responsáveis, pode ser observada na Figura 1.

Modelo	Número de Parâmetros	Empresa / Organização
openchat-3.5-0106	~8B	OpenChat (comunidade OSS)
meta-llama-3.1-8b	~8B	Meta AI
gpt-oss-20b	~20B	OpenAI OSS
Gemini 1.5 Flash	~8B	Google DeepMind

Tabela 1 – Quantidade de parâmetros dos modelos de LLM usados no projeto.

3.5 Avaliação de Modelos de Linguagem de Grande Porte (LLMs)

O sistema RAG foi integrado a três modelos distintos de linguagem de grande porte (LLMs): o *Gemini Pro*, o *GPT-4*, e o *OpenChat*, além de uma quarta abordagem local utilizando o modelo *Ollama (LLaMA 3)*. Cada modelo foi avaliado ao responder um conjunto padronizado de perguntas técnicas formuladas com base nos conteúdos dos documentos utilizados. A avaliação buscou analisar aspectos como precisão, clareza e adequação normativa das respostas geradas.

3.6 Elaboração e Aplicação do Conjunto de Perguntas

Foi elaborado um conjunto de perguntas com base nos temas centrais dos documentos técnicos indexados, como pode ser visto no Quadro 1. As perguntas abordavam tópicos como cálculo de tração, módulo de elasticidade, normas de segurança e parâmetros técnicos típicos de projetos de transformadores. Cada pergunta foi submetida aos modelos selecionados, com as respectivas respostas sendo organizadas em planilhas para posterior análise comparativa.

Legenda	Perguntas
Pergunta 1	Quais são os requisitos para autorização de um trabalhador realizar trabalho em altura?
Pergunta 2	O que a NR35 define sobre o uso do cinturão de segurança tipo paraquedista?
Pergunta 3	Quais devem ser as etapas do planejamento para atividades em altura?
Pergunta 4	Qual a validade da Permissão de Trabalho (PT) segundo a NR35?
Pergunta 5	Como devem ser realizados os treinamentos exigidos pela norma?
Pergunta 6	O que é o Módulo de Young e como ele é aplicado na engenharia?
Pergunta 7	Como o módulo de elasticidade está relacionado à deformação dos materiais?
Pergunta 8	Quais materiais possuem maior módulo de elasticidade?
Pergunta 9	Em que situações é importante considerar o módulo de elasticidade na indústria?
Pergunta 10	O módulo de Young varia de acordo com a temperatura ou outras condições ambientais?
Pergunta 11	Como é definida a força de tração?
Pergunta 12	Qual a equação básica para o cálculo da tração em um corpo?
Pergunta 13	Quais são os exemplos práticos do uso da força de tração?
Pergunta 14	O que diferencia tração de compressão?
Pergunta 15	A tração depende do ângulo da força aplicada sobre o corpo?

Quadro 1 – Demonstra as perguntas direcionadas as LLMs.

Fonte: Elaborado pelo autor, 2025.

3.7 Validação e Avaliação Qualitativa

As respostas geradas pelos modelos foram avaliadas por especialistas com experiência profissionais e acadêmicas na área da engenharia, como visualizado no Quadro 2. A análise considerou critérios como: aderência ao conteúdo normativo, rastreabilidade das respostas (quando possível), clareza da explicação e coerência técnica. Cada resposta foi pontuada em uma escala de 0 a 10, conforme sua adequação, e os resultados foram consolidados em uma tabela comparativa que permitiu inferências sobre a performance de cada modelo frente ao mesmo conjunto de dados e perguntas.

Avaliador	Historico Academico	Experiência profissional
Filipe	Técnico em Eletrotécnica (SENAI), Engenheiro de Controle e Automação (IFMG)	2 anos na área de Ciência e Engenharia de Dados e 2 anos atuando na área de desenvolvimento de software
Avaliador 1	Técnico em Eletrotécnica (SENAI), Engenheiro Mecânico (PUC Minas)	5 anos na área de projetos elétricos
Avaliador 2	Técnico em Segurança do Trabalho (SENAI), Engenheiro Mecânico (PUC Minas)	2 anos na área de engenharia do produto
Avaliador 3	Técnico em Química (CEFET-MG), Engenheiro Metalúrgico (UFMG), Mestrado em Processo de Fabricação com Ênfase em Soldagem (UFMG)	18 anos como Supervisor na área de Engenharia do Produto, 11 anos como Inspetor de Fabricação e 10 anos como Professor Universitário para materias voltadas para processos de fabricação e tecnologia dos materiais.

Quadro 2 – Apresentação dos indivíduos que avaliaram as LLMs.

Fonte: Elaborado pelo autor, 2025.

3.8 Limitações Técnicas e Perspectivas Futuras

Apesar do uso de modelos robustos, o projeto enfrentou limitações em termos de infraestrutura local, especialmente para execução de modelos LLM com maior número de parâmetros. O modelo de 20B, por exemplo, demandou recursos computacionais significativamente superiores. Além disso, algumas respostas apresentaram inconsistências associadas à forma de recuperação ou à ambiguidade das perguntas. Como perspectiva futura, propõe-se o refinamento do retriever com ajuste fino no modelo de embeddings, bem como a ampliação do número de documentos e tipos de normas processadas.

4 PROPOSTA E DESENVOLVIMENTO

4.1 Proposta

A proposta deste trabalho, representada de forma sucinta na Figura 3, consistiu na criação de uma solução computacional baseada na técnica de *Retrieval-Augmented Generation* (RAG), voltada para o suporte a projetistas e engenheiros na consulta de documentos técnicos relacionados à fabricação de transformadores de grande porte. O objetivo central foi estruturar um sistema capaz de processar consultas em linguagem natural e fornecer respostas fundamentadas em trechos de documentos previamente indexados, garantindo rastreabilidade e confiabilidade das informações.

Diferentemente de soluções tradicionais que dependem exclusivamente da memória paramétrica dos modelos de linguagem, a arquitetura RAG permite integrar um mecanismo de recuperação documental a modelos de geração, assegurando maior precisão técnica e mitigando o risco de respostas imprecisas ou inventadas. Além disso, a escolha por rodar o sistema em ambiente local (*on-premises*) visou preservar a segurança dos dados corporativos, eliminando a dependência de APIs externas e reduzindo o risco de vazamento de informações sensíveis.

A proposta contemplou ainda a comparação entre diferentes LLMs — tanto modelos locais, como o *openchat-3.5-0106* e o *meta-llama-3.1-8b*, quanto modelos de maior porte, como o *gpt-oss-20b* e o *Gemini 1.5 Flash* — de forma a avaliar o impacto da quantidade de parâmetros e do ambiente de execução sobre a qualidade das respostas.

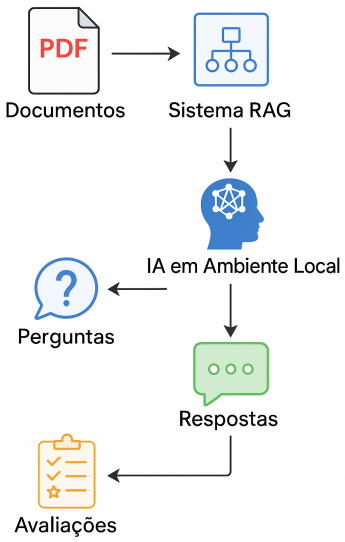

Figura 3 – Fluxograma da proposta de desenvolvimento.

Fonte: Elaborado pelo autor, 2025.

4.2 Desenvolvimento

O desenvolvimento da solução foi conduzido em etapas sucessivas, cada uma correspondendo a um módulo da arquitetura RAG implementada em ambiente Jupyter Notebook. O fluxo geral contemplou desde o processamento inicial dos documentos até a avaliação comparativa dos modelos de linguagem.

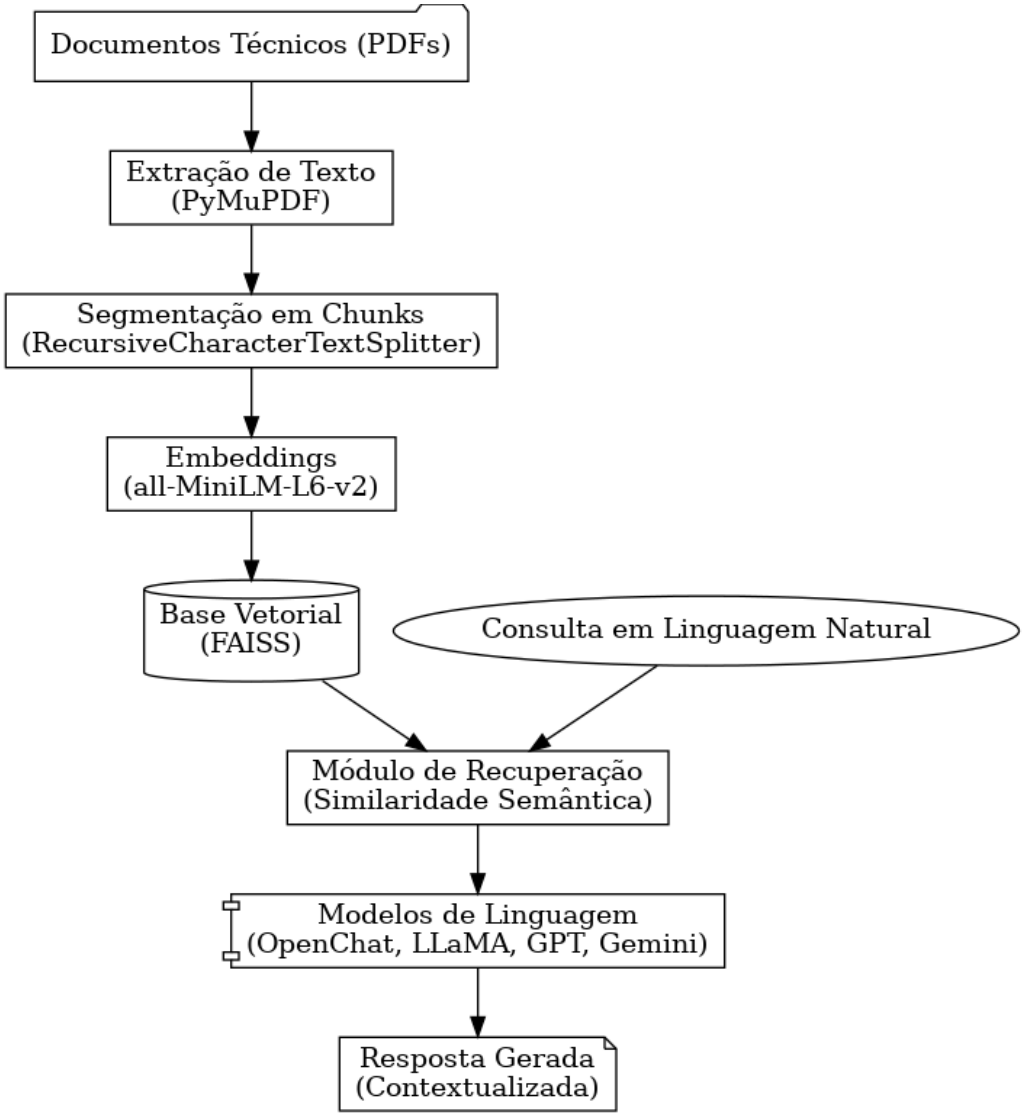

Figura 4 – Fluxograma da arquitetura RAG implementada no projeto.

Fonte: Elaborado pelo autor, 2025.

A Figura 4 apresenta o fluxograma da arquitetura RAG implementada no projeto. O processo inicia-se com os **documentos técnicos em PDF**, que representam as Instruções Técnicas (ITs) e demais normas utilizadas no contexto da engenharia de transformadores. Esses documentos passam por um processo de **extração textual**, realizado com a biblioteca *PyMuPDF*, convertendo o conteúdo em texto bruto processável.

Na etapa seguinte, aplica-se a **segmentação em *chunks***, utilizando a ferramenta

RecursiveCharacterTextSplitter, a qual divide o texto em trechos menores com sobreposição. Essa estratégia tem como objetivo preservar a continuidade semântica, evitando perda de informações contextuais importantes durante a recuperação.

Os trechos segmentados são então convertidos em representações numéricas por meio da geração de **embeddings**, realizada pelo modelo *all-MiniLM-L6-v2*. Esses vetores são indexados e armazenados em uma **base vetorial FAISS**, que possibilita a busca eficiente de trechos semanticamente similares.

Quando o usuário submete uma **consulta em linguagem natural**, esta é processada pelo **módulo de recuperação**, que compara semanticamente a pergunta com os vetores armazenados na base. Os trechos mais relevantes são então enviados como contexto adicional para os **modelos de linguagem de grande porte (LLMs)**, como *OpenChat*, *LLaMA*, *GPT* e *Gemini*.

Por fim, o LLM gera uma **resposta contextualizada**, fundamentada nas informações extraídas dos documentos técnicos. Esse fluxo garante maior rastreabilidade, precisão técnica e reduz significativamente a ocorrência de respostas inventadas, características que justificam a adoção da arquitetura RAG neste trabalho.

5 ANÁLISE DOS RESULTADOS E IMPACTOS

O desenvolvimento deste trabalho buscou demonstrar a viabilidade de utilização de Modelos de Linguagem de Grande Porte (LLMs) em ambiente local, de modo a eliminar riscos de vazamento de informações sensíveis. Para isso, estruturou-se uma arquitetura RAG totalmente executada em servidor local, sem dependência de serviços externos.

Entretanto, durante a execução, observou-se que a máquina disponível apresentava limitações de hardware, o que impactou o desempenho de modelos mais robustos. Os LLMs utilizados neste estudo variaram entre 8B e 20B parâmetros, demandando elevado processamento e memória. Apesar disso, em um cenário empresarial, onde há possibilidade de maior investimento em infraestrutura computacional, a adoção de modelos com maior número de parâmetros seria plenamente viável e potencialmente mais eficiente.

5.1 Avaliação das Respostas dos Modelos

A todos os avaliadores foi entregue um formulário no formato ilustrado na Tabela 2. Nesse documento, a primeira coluna apresentava todas as perguntas elaboradas, seguidas por quatro colunas contendo as respostas fornecidas por cada um dos modelos avaliados. Além disso, foram disponibilizadas quatro colunas adicionais em branco, destinadas ao registro das notas atribuídas pelos avaliadores a cada resposta correspondente.

Pergunta	IA-1	IA-2	IA-3	IA-4	Nota-IA-1	Nota-IA-2	Nota-IA-3	Nota-IA-4
Pergunta 1	Respostas-IA-1	Respostas-IA-2	Respostas-IA-3	Respostas-IA-4	x	x	x	x
Pergunta 2	Respostas-IA-1	Respostas-IA-2	Respostas-IA-3	Respostas-IA-4	x	x	x	x
Pergunta 3	Respostas-IA-1	Respostas-IA-2	Respostas-IA-3	Respostas-IA-4	x	x	x	x
Pergunta 4	Respostas-IA-1	Respostas-IA-2	Respostas-IA-3	Respostas-IA-4	x	x	x	x
Pergunta 5	Respostas-IA-1	Respostas-IA-2	Respostas-IA-3	Respostas-IA-4	x	x	x	x

Tabela 2 – Exemplar do documento entregue aos avaliadores.

Obs: Tabela com conteúdo ilustrativo para fins didáticos.

Como critérios de avaliação, foram adotadas as referências apresentadas no Quadro 3. Esses parâmetros forneceram aos avaliadores uma base padronizada para a análise, permitindo atribuir as notas de forma mais consistente e precisa a cada resposta dos modelos.

Faixa	Significado
0–3	Resposta incorreta, irrelevante ou genérica
4–6	Responde parcialmente, mas sem profundidade técnica
7–8	Resposta correta, com clareza e técnica moderada
9–10	Resposta precisa, clara, e referenciada com conteúdo técnico relevante

Quadro 3 – Exemplar do documento entregue aos avaliadores.

A Figura 5 apresenta um gráfico com as notas atribuídas pelo autor como referência inicial, servindo de base para comparação com as avaliações posteriores.

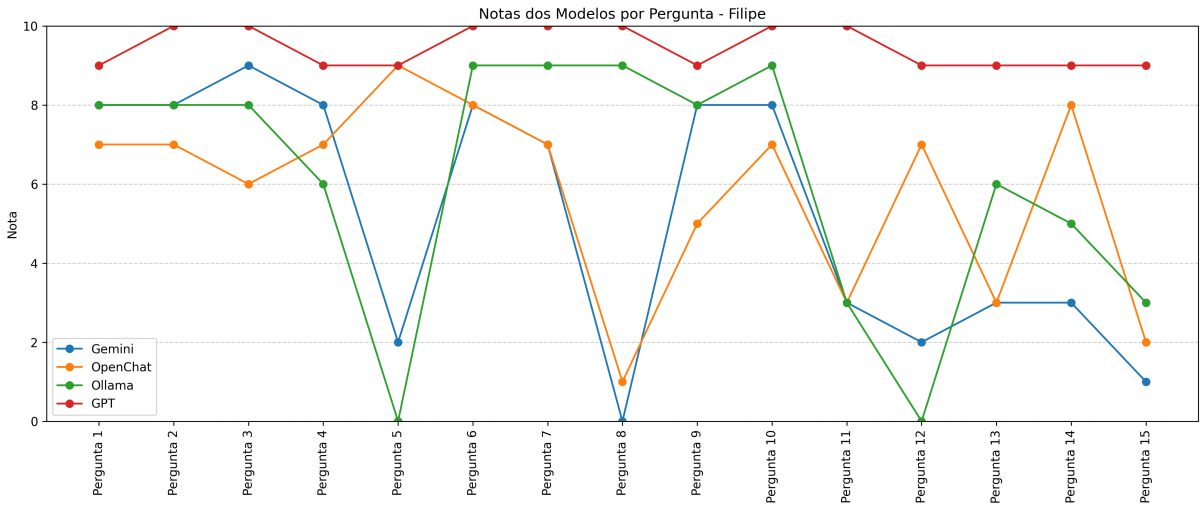

Figura 5 – Notas atribuídas pelo autor (base comparativa).

Fonte: Elaborado pelo autor, 2025.

A Figura 6 mostra as notas atribuídas pelo Avaliador 1, profissional com formação em Engenharia Mecânica e Técnico em Eletrotécnica.

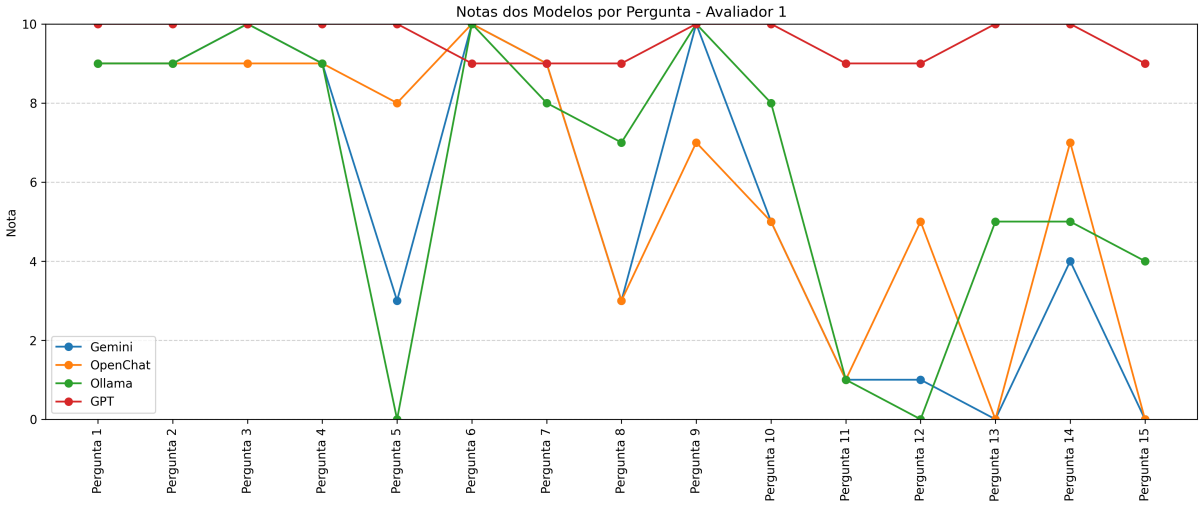

Figura 6 – Notas atribuídas pelo Avaliador 1.

Fonte: Elaborado pelo autor, 2025.

A Figura 7 apresenta as avaliações realizadas pelo Avaliador 2, Engenheiro Mecânico e Técnico em Segurança do Trabalho.

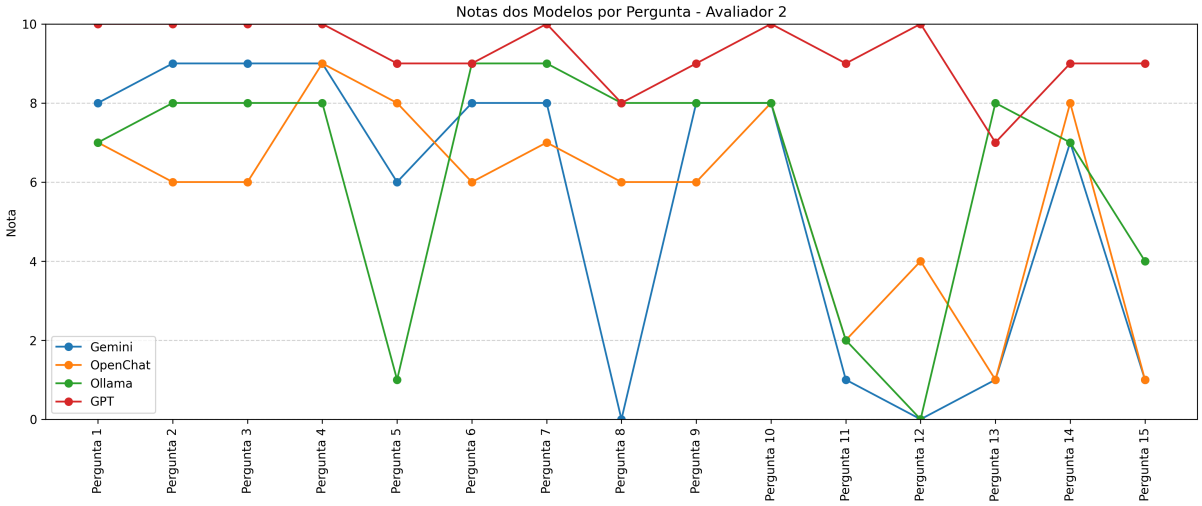

Figura 7 – Notas atribuídas pelo Avaliador 2.

Fonte: Elaborado pelo autor, 2025.

A Figura 8 apresenta as notas do Avaliador 3, Mestre em Processos de Fabricação com ênfase em Soldagem, Engenheiro Metalúrgico e Técnico em Química.

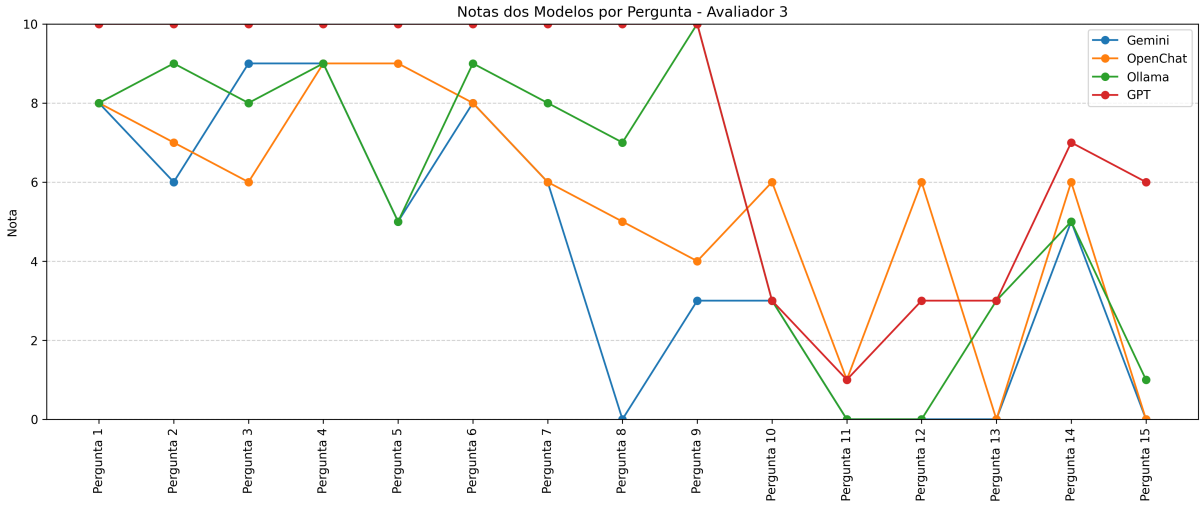

Figura 8 – Notas atribuídas pelo Avaliador 3.

Fonte: Elaborado pelo autor, 2025.

Além das avaliações humanas, também foram consideradas análises automáticas realizadas pelas próprias IAs Gemini e ChatGPT, conforme ilustrado nas Figuras 9 e 10. Contudo, é importante destacar que ambos os modelos tinham acesso à informação sobre a autoria de cada resposta, o que pode ter introduzido um viés nos resultados obtidos.

Para a execução dessas análises, foi utilizado o seguinte *prompt* de avaliação:

> "Avalie a resposta de alguns modelos de LLM com base no critério do documento `Escala_Pontuacao.xlsx` (conteúdo equivalente ao Quadro 3). No documento `Respostas_Modelos_RAG.xlsx` (conteúdo equivalente à Tabela 2), preencha as colunas: `Nota_Gemini`, `Nota_OpenChat`, `Nota_Ollama`, `Nota_GPT` e, abaixo, na mesma linha das respostas, coloque o valor atribuído a cada uma delas referente à pergunta da linha."

A Figura 9 apresenta as notas atribuídas pelo modelo *Gemini 1.5 Flash*, disponibilizado gratuitamente pela Google.

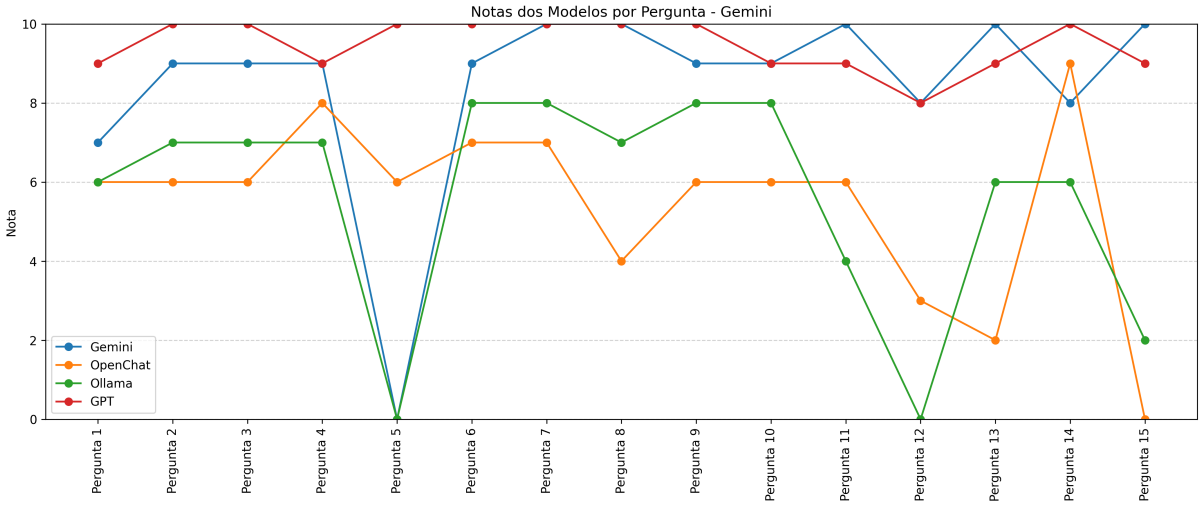

Figura 9 – Notas atribuídas pela IA Gemini.

Fonte: Elaborado pelo autor, 2025.

Já a Figura 10 mostra as notas geradas pelo modelo *ChatGPT 5*, acessado por meio de assinatura e fornecido pela OpenAI.

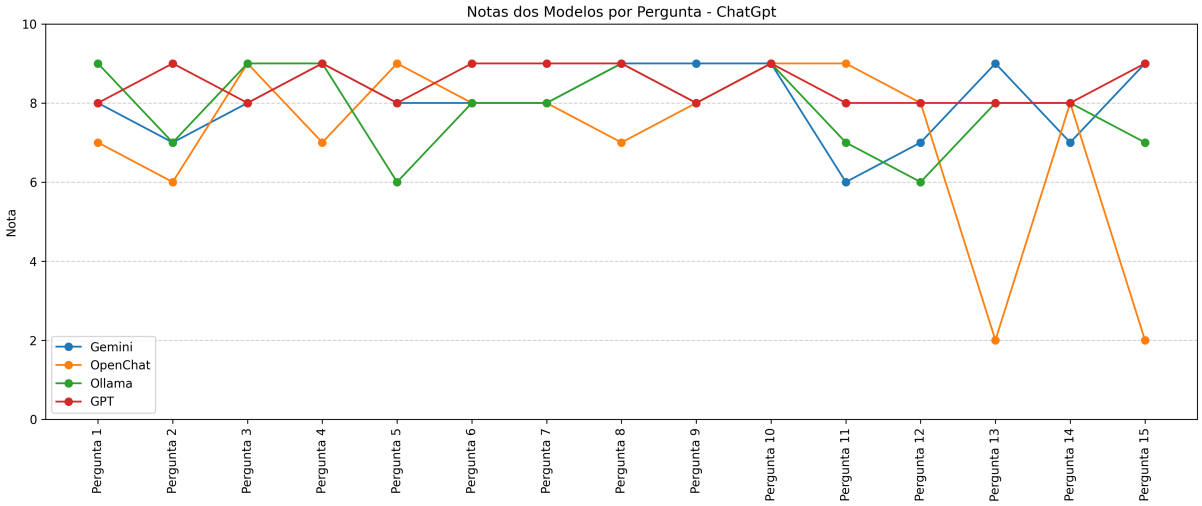

Figura 10 – Notas atribuídas pela IA ChatGPT.

Fonte: Elaborado pelo autor, 2025.

Por fim, a Figura 11 sintetiza a pontuação total de cada modelo, considerando a soma das notas atribuídas por todos os avaliadores.

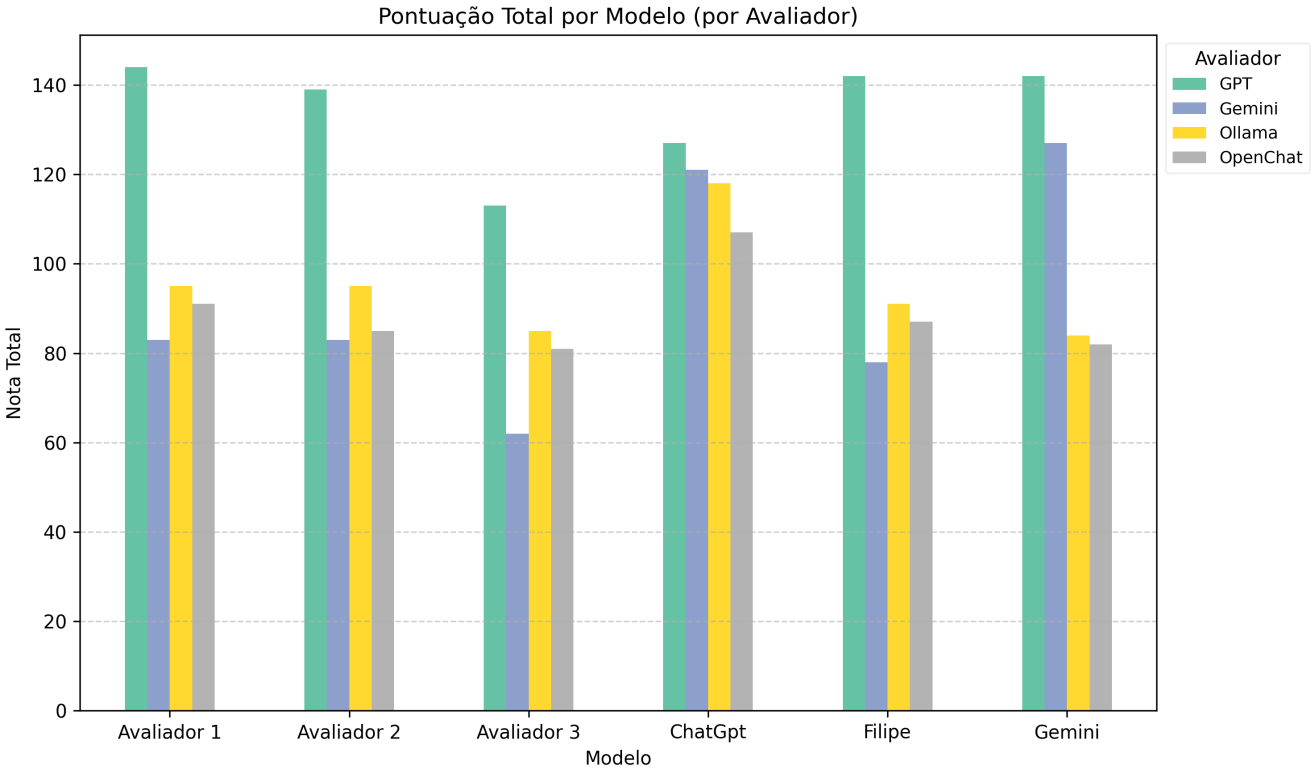

Figura 11 – Pontuação total consolidada por modelo.

Fonte: Elaborado pelo autor, 2025.

5.2 Análise Comparativa dos Modelos

A partir das avaliações realizadas, é possível destacar os seguintes pontos sobre o desempenho de cada LLM:

- **GPT-OSS-20B**: Apresentou desempenho mais consistente entre os avaliadores, atingindo frequentemente notas próximas de 9 e 10. Sua maior quantidade de parâmetros refletiu em respostas mais completas e tecnicamente fundamentadas, embora tenha exigido maior capacidade computacional.

- **Gemini 1.5 Flash**: Demonstrou desempenho sólido, especialmente em perguntas de maior contextualização, com notas médias entre 8 e 9. Destacou-se pela rapidez nas respostas, mesmo em ambiente de API, mas apresentou variações de consistência em alguns casos específicos.

- **Meta-LLaMA 3.1-8B (via Ollama)**: Teve desempenho intermediário, alcançando notas entre 6 e 8 na maior parte das avaliações. Apesar de menos robusto que os modelos maiores, apresentou boa relação entre custo computacional e qualidade de resposta.

- **OpenChat 3.5**: Foi o modelo que mais oscilou entre as perguntas, com notas variando de muito baixas até 9 em alguns casos. Apesar de ter apresentado limitações em consistência, demonstrou potencial em perguntas mais diretas, evidenciando-se como uma alternativa viável em cenários de menor exigência técnica.

De forma geral, os resultados confirmam que modelos com maior número de parâmetros (como o GPT-OSS-20B) fornecem respostas mais confiáveis e estáveis, mas com elevado custo computacional. Modelos menores (como LLaMA 3.1-8B e OpenChat) oferecem uma alternativa prática em ambientes com restrição de recursos, ainda que com perda de consistência. O Gemini 1.5 Flash apresentou-se como uma solução equilibrada, conciliando boa performance com eficiência operacional.

6 CONCLUSÕES

O presente trabalho teve como objetivo principal investigar a viabilidade da utilização de Modelos de Linguagem de Grande Porte (LLMs) em conjunto com a técnica de *Retrieval-Augmented Generation* (RAG), operando em ambiente local (*on-premises*), como alternativa segura para acesso a informações técnicas sensíveis. A proposta contemplou a criação de uma arquitetura funcional, a construção de uma base documental simulando Instruções Técnicas (ITs) e a avaliação comparativa do desempenho de diferentes modelos.

A implementação demonstrou que é tecnicamente possível estruturar um sistema de consulta baseado em RAG inteiramente em servidor local, garantindo maior controle sobre os dados e eliminando riscos de vazamento de informações para redes externas. Embora tenham sido encontradas limitações relacionadas à infraestrutura disponível, sobretudo no que diz respeito à execução de modelos com 20 bilhões de parâmetros, verificou-se que, em ambientes corporativos com maior capacidade computacional, tais restrições podem ser superadas.

Os resultados obtidos evidenciaram diferenças significativas entre os modelos testados. O GPT-OSS-20B destacou-se pela consistência e qualidade das respostas, ainda que demandando recursos computacionais elevados. O Gemini 1.5 Flash apresentou-se como uma solução equilibrada, com boa precisão e eficiência operacional. O Meta-LLaMA 3.1-8B demonstrou desempenho intermediário, adequado para cenários com restrição de hardware, enquanto o OpenChat 3.5 revelou maior instabilidade, mas se mostrou promissor em perguntas mais objetivas.

Em termos práticos, a pesquisa validou a hipótese de que sistemas RAG locais podem ser aplicados de forma efetiva em ambientes industriais e corporativos, oferecendo respostas contextualizadas e tecnicamente fundamentadas sem comprometer a segurança das informações. A análise comparativa também reforçou a importância de alinhar a escolha do modelo de linguagem às condições de infraestrutura disponíveis e às necessidades do usuário final.

Trabalhos Futuros

Como continuidade deste estudo, sugerem-se os seguintes aprimoramentos:

- Expansão da base documental, contemplando um número maior de normas, manuais e instruções técnicas específicas da área industrial.
- Investigação de técnicas avançadas de recuperação, como reranqueamento semântico e geração de documentos hipotéticos, a fim de aumentar a precisão e a contextualização

das respostas.

- Avaliação do impacto da utilização de LLMs mais recentes e com diferentes arquiteturas, explorando o equilíbrio entre desempenho, custo computacional e confiabilidade.
- Desenvolvimento de uma interface gráfica intuitiva, que facilite a adoção do sistema por profissionais de diferentes níveis de especialização.

Em síntese, este trabalho demonstrou a viabilidade da integração entre RAG e LLMs em ambiente local como uma solução prática, segura e promissora para o acesso a informações técnicas em contextos industriais, abrindo espaço para novas aplicações e pesquisas que explorem o potencial da inteligência artificial em domínios críticos e altamente especializados.

REFERÊNCIAS

ARSLAN, M. *et al.* A survey on rag with llms. **Procedia Computer Science**, Amsterdam, v. 246, p. 3781–3790, 2024. Disponível em: https://doi.org/10.1016/j.procs.2024.09.178. Acesso em: 2025-04-15.

BORA, A.; CUAYáHUITL, H. Systematic analysis of retrieval-augmented generation-based llms for medical chatbot applications. **Machine Learning and Knowledge Extraction**, Basel, v. 6, n. 4, p. 2355–2374, 2024. Disponível em: https://doi.org/10.3390/make6040116. Acesso em: 2025-04-15.

CHO, S.; PARK, J.; UM, J. Development of fine-tuned retrieval augmented language model specialized to manual books on machine tools. **IFAC-PapersOnLine**, v. 58, n. 19, p. 187–192, 2024. Disponível em: https://doi.org/10.1016/j.ifacol.2024.09.157. Acesso em: 2025-04-15.

GARCIA, P.; SILVA, A.; OLIVEIRA, L. **Chatbot Inteligente para Acesso a Documentos Acadêmicos**. 2025. Dissertação (Mestrado) — Instituto Federal de Educação, Ciência e Tecnologia, 2025. Trabalho de Conclusão de Curso.

GARCíA, S. D. **Desenvolvemento dun sistema de consulta en linguaxe natural para a recuperación de documentación empresarial mediante IA xenerativa e técnicas RAG**. 2025. Dissertação (Mestrado) — Universidade da Coruña, 2025. Traballo Fin de Máster.

JAGUáN, E. A. R. **Arquitectura Multiagente para Modelos RAG en Entornos de Tráfico Aéreo**. 2025. Dissertação (Mestrado) — Universidad Politécnica de Madrid, 2025. Trabajo Final de Grado.

KURATOMI, R.; KAWAKUBO, A. L.; FALCãO, T. C. A. Inteligência artificial e recuperação de informações jurídico-institucionais: aplicação do rag na universidade de são paulo. *In*: SBC. **Anais do Congresso da Sociedade Brasileira de Computação**. São Paulo, 2025.

NAVEED, H. *et al.* A comprehensive overview of large language models. **arXiv preprint**, arXiv:2307.06435v10, 2024. Disponível em: https://arxiv.org/abs/2307.06435. Acesso em: 2025-04-15.

PARANHOS, S. L.; SOUZA, J. C.; ALMEIDA, R. Avaliação do impacto de diferentes padrões arquiteturais rag em domínios jurídicos. *In*: **Anais do Congresso da Sociedade Brasileira de Computação**. [*S.l.: s.n.*], 2024. Disponível em: https://doi.org/10.5753/cbie.2024.32216. Acesso em: 2025-04-15.

SILVA, V. F. d. **Assistente Virtual para Coordenador de Curso: Aplicação de LLMs à Gestão de Perguntas Frequentes em um Curso de Graduação**. 2025. Dissertação (Mestrado) — Universidade Federal de Santa Catarina, 2025. Trabalho de Conclusão de Curso.

SIRIWARDHANA, S. *et al.* Improving the domain adaptation of retrieval-augmented generation (rag) models for open domain question answering. **Transactions of the**

Association for Computational Linguistics, v. 11, p. 1–17, 2023. Disponível em: https://doi.org/10.1162/tacl_a_00530. Acesso em: 2025-04-15.

YU, Y.; ZHANG, D.; CHU, C. A survey of multimodal large language models. *In*: **31st Annual Meeting of the Association for Natural Language Processing**. [*S.l.: s.n.*], 2025. Disponível em: https://mm-llms.github.io. Acesso em: 2025-04-15.