UNIVERSIDADE FEDERAL DO PARÁ

INSTITUTO DE TECNOLOGIA

FACULDADE DE ENGENHARIAS ELÉTRICA E BIOMÉDICA

ANDRÉ OLIVEIRA CARVALHO DA SILVA

CHATBOT INTELIGENTE PARA AUXÍLIO À TOMADA DE DECISÃO EM EMPRESA DO SETOR DE ENERGIA

Belém – PA

2025

ANDRÉ OLIVEIRA CARVALHO DA SILVA

CHATBOT INTELIGENTE PARA AUXÍLIO À TOMADA DE DECISÃO EM EMPRESA DO SETOR DE ENERGIA

Trabalho de Conclusão de Curso apresentado à Faculdade de Engenharias Elétrica e Biomédica do Instituto de Tecnologia da Universidade Federal do Pará, como requisito para a obtenção do Grau de Bacharel em Engenharia Elétrica.

Orientadora: MSc. Elen Priscila de Souza Lobato

Coorientador: Prof. Dr. Wellington da Silva Fonseca

Belém – PA

2025

Dados Internacionais de Catalogação na Publicação (CIP) de acordo com ISBD
Sistema de Bibliotecas da Universidade Federal do Pará
Gerada automaticamente pelo módulo Ficat, mediante os dados fornecidos pelo(a) autor(a)

D111c Da Silva, André Oliveira Carvalho.
CHATBOT INTELIGENTE PARA AUXÍLIO À TOMADA DE DECISÃO EM EMPRESA DO SETOR DE ENERGIA / André Oliveira Carvalho Da Silva. — 2025.
43 f. : il. color.

Orientador(a): Profª. MSc. Elen Priscila de Souza Lobato
Coorientador(a): Prof. Dr. Welington da Silva Fonseca
Trabalho de Curso (Graduação) - Universidade Federal do Pará, Instituto de Tecnologia, Faculdade de Engenharia Elétrica, Belém, 2025.

1. Chatbot. 2. Setor Elétrico. 3. Recuperação Aumentada por Geração. 4. Modelos de Linguagem de Larga Escala. 5. Tomada de Decisão Operacional. I. Título.

CDD 006.3

ANDRÉ OLIVEIRA CARVALHO DA SILVA

CHATBOT INTELIGENTE PARA AUXÍLIO À TOMADA DE DECISÃO EM EMPRESA DO SETOR DE ENERGIA

Trabalho de Conclusão de Curso apresentado à Faculdade de Engenharias Elétrica e Biomédica do Instituto de Tecnologia da Universidade Federal do Pará, como requisito para a obtenção do Grau de Bacharel em Engenharia Elétrica.

Data de aprovação:

Conceito:

BANCA EXAMINADORA:

_____

MSc. Elen Priscila de Souza Lobato
Orientadora – CEAMAZON / UFPA

_____

Prof. Dr. Welington da Silva Fonseca
Coorientador – Faculdade de Engenharia Elétrica e Biomédica - FEEB / ITEC / UFPA

_____

Prof. Dr. Miércio Cardoso de Alcântara Neto
Avaliador Interno – FEEB / ITEC / UFPA

_____

Eng. Eletric. Lusiane Pereira Fonseca
Avaliador Externo – Grupo Equatorial Energia/PA

VISTO:

_____

Prof. Dr.ª Carminda Célia Moura de Moura Carvalho
Diretora da Faculdade de Engenharia Elétrica e Biomédica – FEEB / UFPA

AGRADECIMENTOS

Gostaria de agradecer primeiramente a Deus por me conceder forças durante minha jornada acadêmica e pela oportunidade de viver esta experiência.

Aos meus pais, Ewerton e Ângela, pelo apoio incansável de toda uma vida. Obrigado por sempre me encorajarem a buscar o meu melhor, tanto no âmbito acadêmico quanto no pessoal. Vocês são meus maiores exemplos de dedicação, força e caráter. Levo comigo cada ensinamento que me passaram e sou profundamente grato por tudo que fizeram e continuam fazendo por mim.

Ao meu irmão, Pedro, pelo incentivo constante ao longo de toda a minha trajetória na faculdade, desde a escolha do curso até os ensinamentos e a ajuda essencial nos primeiros semestres. Agradeço pela motivação contínua, que sempre me impulsionou a seguir em frente, e por acreditar em mim em todas as etapas desse caminho.

À minha namorada, Emanuele, por sempre acreditar em mim e me lembrar, nos momentos difíceis, de seguir em frente. Obrigado pelo apoio constante ao longo desses anos e pelo incentivo diário que fez toda a diferença nessa jornada. Nada disso teria sido possível sem você ao meu lado. Sou imensamente grato por tê-la comigo durante essa fase tão importante da minha vida.

Aos meus amigos, que ao longo desse período contribuíram para tornar a jornada mais leve, seja nas conversas, nos estudos ou na convivência do dia a dia, e cujas amizades tiveram papel importante durante essa etapa.

Ao Prof. Dr. Wellington da Silva Fonseca e à MSc. Elen Priscila de Souza Lobato, pela orientação, pelo apoio e pelas contribuições essenciais para a realização deste trabalho. Agradeço também as contribuições dos membros da banca, Prof. Dr. Miércio Alcântara e a Engenheira Eletricista Lusiane Fonseca.

E, por fim, à Universidade Federal do Pará e à Equatorial Energia, pelo apoio e pelas oportunidades que contribuíram para o desenvolvimento deste trabalho.

Aos meus pais e ao meu irmão, pelo suporte constante e pela força que me ofereceram ao longo de toda a minha trajetória acadêmica.

INFORMATIVO

Neste documento estão presentes o artigo publicado na *International Conference on Industry Applications* (INDUCON 2025), a apresentação da defesa do Trabalho de Conclusão de Curso, assim como sua respectiva comprovação de publicação, utilizados para o crédito na disciplina de Trabalho de Conclusão de Curso, de acordo com a Instrução Normativa Nº 01/2022 da Pró-Reitora de Ensino de Graduação (PROEG/UFPA).

RESUMO

Este trabalho apresenta o desenvolvimento de um chatbot inteligente voltado para auxiliar a tomada de decisão operacional em uma concessionária de energia elétrica da região Norte do Brasil. A aplicação utiliza Modelos de Linguagem de Larga Escala (LLMs), *embeddings* semânticos e a técnica de Recuperação Aumentada por Geração (RAG) para oferecer respostas precisas, contextualizadas e multimodais, a partir de uma base de dados textual e visual indexada. A arquitetura do sistema foi estruturada com o uso da biblioteca *LangGraph*, que permite o controle dinâmico do fluxo conversacional, e do *Milvus*, um sistema de gerenciamento de dados vetoriais. A interface desenvolvida com *HTML, CSS e JavaScript* proporciona uma experiência interativa e acessível ao usuário. Testes demonstraram precisão nas respostas e tempo médio de resposta inferior a dois segundos, evidenciando o potencial do chatbot como ferramenta estratégica para aumento da eficiência operacional no setor elétrico. O estudo destaca a viabilidade técnica e o impacto positivo da aplicação de IA generativa em ambientes industriais críticos.

Palavras-chave: Chatbot; Setor Elétrico; Recuperação Aumentada por Geração; Modelos de Linguagem de Larga Escala; Tomada de Decisão Operacional.

ABSTRACT

This work presents the development of an intelligent chatbot designed to assist operational decision-making in an electric power utility company located in the Northern region of Brazil. The application employs Large Language Models (LLMs), semantic embeddings, and the Retrieval-Augmented Generation (RAG) technique to provide accurate, contextualized, and multimodal responses based on an indexed textual and visual database. The system architecture was structured using the LangGraph library, which enables dynamic control of the conversational flow, and Milvus, a vector data management system. The interface, developed with HTML, CSS, and JavaScript, provides an interactive and user-friendly experience. Tests demonstrated accuracy in the responses and an average response time of less than two seconds, highlighting the chatbot's potential as a strategic tool for enhancing operational efficiency in the electrical sector. The study underscores the technical feasibility and positive impact of applying generative AI in critical industrial environments.

Keywords: Chatbot; Electric Power Sector; Retrieval-Augmented Generation; Large Language Models; Operational Decision-Making.

FIGURAS

TABELAS

LISTA DE ABREVIATURAS E SIGLAS

AI	*Artificial Intelligence*
CSS	*Cascading Style Sheets*
HTML	*HyperText Markup Language*
IA	Inteligência Artificial
JSON	*JavaScript Object Notation*
KV	Quilovolts
LLM / LLMs	*Large Language Model / Large Language Models*
LT	Linha de Transmissão
PDF	*Portable Document Format*
PLN	Processamento de Linguagem Natural
RAG	*Retrieval-Augmented Generation*
SQL	*Structured Query Language*
UFPA	Universidade Federal do Pará
LGPD	Lei Geral de Proteção de Dados Pessoais

SUMÁRIO

Chatbot Inteligente Para Auxílio À Tomada De Decisão Em Empresa Do Setor De Energia

A.O.C da Silva, E.D. Silva, E.P. de S. Lobato, W. da S. Fonseca, L.P. Fonseca.

Abstract: *This work presents the development of an intelligent chatbot designed to assist operational decision-making in an electric power utility company located in the Northern region of Brazil. The application employs Large Language Models (LLMs), semantic embeddings, and the Retrieval-Augmented Generation (RAG) technique to provide accurate, contextualized, and multimodal responses based on an indexed textual and visual database. The system architecture was structured using the LangGraph library, which enables dynamic control of the conversational flow, and Milvus, a vector data management system. The interface, developed with HTML, CSS, and JavaScript, provides an interactive and user-friendly experience. Tests demonstrated accuracy in the responses and an average response time of less than two seconds, highlighting the chatbot's potential as a strategic tool for enhancing operational efficiency in the electrical sector. The study underscores the technical feasibility and positive impact of applying generative AI in critical industrial environments.*

Resumo: Este trabalho apresenta o desenvolvimento de um chatbot inteligente voltado para auxiliar a tomada de decisão operacional em uma concessionária de energia elétrica da região Norte do Brasil. A aplicação utiliza Modelos de Linguagem de Larga Escala (LLMs), *embeddings* semânticos e a técnica de Recuperação Aumentada por Geração (RAG) para oferecer respostas precisas, contextualizadas e multimodais, a partir de uma base de dados textual e visual indexada. A arquitetura do sistema foi estruturada com o uso da biblioteca *LangGraph*, que permite o controle dinâmico do fluxo conversacional, e do *Milvus*, um sistema de gerenciamento de dados vetoriais. A interface desenvolvida com *HTML, CSS e JavaScript* proporciona uma experiência interativa e acessível ao usuário. Testes demonstraram precisão nas respostas e tempo médio de resposta inferior a dois segundos, evidenciando o potencial do chatbot como ferramenta estratégica para aumento da eficiência operacional no setor elétrico. O estudo destaca a viabilidade técnica e o impacto positivo da aplicação de IA generativa em ambientes industriais críticos.

Keywords: *Chatbot, Electric Power Sector, Retrieval-Augmented Generation, Large Language Models, Operational Decision-Making.*

Palavras-chave: Chatbot, Setor Elétrico, Recuperação Aumentada por Geração, Modelos de Linguagem de Larga Escala, Tomada de Decisão Operacional.

1. INTRODUÇÃO

O acesso à informação desempenha um papel central na sociedade, evoluindo desde a leitura de jornais impressos até a utilização de ferramentas baseadas em Inteligência Artificial (IA) (DELBIANCO; VALENTIM, 2022). Esse processo de transformação reflete nos avanços tecnológicos e nas mudanças nos modos de produção, circulação e consumo da informação. No passado, o acesso era limitado por fatores como tempo, espaço e meios físicos de distribuição (HERNANDEZ; TOLEDO, 2021). Com o advento da internet e, posteriormente, das tecnologias digitais avançadas, como o armazenamento em nuvem, os mecanismos de busca e os sistemas inteligentes, a informação tornou-se mais acessível, dinâmica e descentralizada (GAÄL; MARTINS, 2022).

Atualmente, vivencia-se uma era marcada pela velocidade com que os dados são gerados e compartilhados, exigindo novas habilidades cognitivas e digitais por parte dos indivíduos e das organizações (PERFETTO; REIS; PALETTA, 2023). O conhecimento, que antes era acumulado em bibliotecas, arquivos físicos e instituições formais, passou a circular por redes interconectadas e plataformas digitais que operam em tempo real (ALBERTIN; ALBERTIN, 2021). Nesse novo cenário digital, a informação deixa de ser apenas um suporte para decisões estratégicas e passa a ocupar uma posição central como recurso de alto valor para a inovação, a competitividade e o desenvolvimento social.

Em meio a essa transformação, ferramentas baseadas em Inteligência Artificial (IA), como sistemas de recomendação, algoritmos preditivos e assistentes virtuais, tornam-se essenciais para lidar com o volume e a complexidade dos dados contemporâneos (MORAIS *et al*., 2020). Essas tecnologias facilitam o acesso à informação, otimizando sua filtragem, análise e personalização e promovem uma interação mais eficiente e alinhada às necessidades específicas dos usuários. Essa capacidade de adaptação, orientada por dados, exemplifica uma das principais características da chamada Indústria 4.0 (OLIVEIRA, 2023).

A Indústria 4.0 é marcada pela integração de tecnologias digitais aos processos produtivos, possibilitando automação inteligente, conectividade em tempo real e análise massiva de dados (ANTÔNIO *et al*., 2025). Nesse contexto, o uso de sistemas

inteligentes favorece a tomada de decisões mais ágil, precisa e contextualizada, elevando os níveis de eficiência e competitividade organizacional (COSTA NETO; CAMPOS, 2023).

As organizações, diante desse panorama, deixam de operar com sistemas exclusivamente executores e passam a adotar soluções tecnológicas capazes de interagir, interpretar e responder de forma mais humanizada (FREIRE, 2024). Entre essas tecnologias, os *chatbots* têm ganhado destaque por sua capacidade de intermediar a interação entre usuários e sistemas computacionais, utilizando linguagem natural (PALLIN *et al*., 2024). Os *chatbots* são programas baseados em regras pré-definidas e aprendizado de máquina, que permitem o acesso automatizado a informações armazenadas em nuvem ou em documentos corporativos (MENEGUELLO *et al*., 2023). Sua implementação promove ganhos significativos na gestão da informação, reduzindo a sobrecarga de atendimentos manuais, otimizando o tempo de resposta e ampliando o acesso a dados estratégicos de forma segura e eficiente.

As aplicações dos *chatbots* são inúmeras e vêm se expandindo progressivamente em diversos setores, com destaque para o ambiente empresarial (LUGLI; LUCCA FILHO, 2020). No contexto corporativo, essas ferramentas têm sido utilizadas tanto no relacionamento com o cliente quanto no suporte interno às equipes e processos organizacionais (MARTINS; PAIVA, 2025).

Por meio da integração com repositórios digitais, sistemas internos e plataformas em nuvem, os *chatbots* permitem que colaboradores acessem conteúdos institucionais, normativas internas, procedimentos operacionais, históricos de atendimento, relatórios técnicos, entre outros documentos relevantes, utilizando linguagem natural e comandos simples (LARUSSA FILHO *et al*., 2022). Essa funcionalidade reduz significativamente o tempo despendido na busca manual por informações, ao mesmo tempo em que minimiza erros e aumenta a eficiência operacional.

Dentro deste contexto, este estudo tem como objetivo apresentar o desenvolvimento e a aplicação de um *chatbot* voltado para uma concessionária de distribuição de energia da região Norte. O foco principal é proporcionar aos técnicos controladores um acesso mais rápido, direto e eficiente a informações operacionais

relevantes, permitindo que consultas sejam feitas de forma ágil e concisa. Dessa maneira, busca-se otimizar os processos de tomada de decisão, especialmente em situações que demandam respostas imediatas e precisão no manejo dos dados técnicos e administrativos.

Para estruturar o desenvolvimento do *chatbot*, este trabalho foi organizado da seguinte forma: a Seção 2 apresenta os trabalhos relacionados, descrevendo as principais abordagens existentes. A Seção 3 descreve os conceitos fundamentais e as bibliotecas utilizadas no projeto. Em seguida, a Seção 4 discute os resultados obtidos, incluindo a interface desenvolvida e o desempenho do sistema. Por fim, a Seção 5 traz as considerações finais, sintetizando os principais achados e contribuições deste trabalho.

2. DESENVOLVIMENTO DE CHATBOT

Estudos presentes na literatura destacam o desenvolvimento de *chatbots* integrados a outras interfaces tecnológicas, com o intuito de facilitar o acesso à informação. Um exemplo disso é o projeto descrito por Parkar *et al*. (2021), no qual foi desenvolvido um *chatbot* baseado na *web*, utilizando bibliotecas de Processamento de Linguagem Natural (PLN) e Aprendizado de Máquina para interagir com os usuários de forma automatizada.

Considerando a constante atualização do currículo acadêmico, os autores implementaram um sistema que permite a modificação dinâmica do banco de dados por meio de uma interface desenvolvida em HTML (*HyperText Markup Language*) e PHP (*Hypertext Preprocessor*), possibilitando a atualização periódica das informações. A proposta do sistema visa melhorar a eficiência e a autonomia dos estudantes universitários, fornecendo dados essenciais diretamente da instituição de ensino.

Em estudo de Lee *et al*. (2024) desenvolvido em parceria com a empresa canadense *OptiMicro Technologies Inc*., foi apresentado um *chatbots* baseado em *Large Language Models* (LLMs) com capacidade de *Retrieval-Augmented Generation* (RAG), voltado para aprimorar o serviço de suporte técnico. O sistema foi construído utilizando a plataforma *Flowise AI* e testado com dados reais da empresa. A principal vantagem da abordagem RAG é sua habilidade de acessar dinamicamente bases de

conhecimento externas, o que permite fornecer respostas atualizadas e sensíveis ao contexto, reduzindo o risco de alucinações típicas em modelos LLM tradicionais. Além disso, a RAG dispensa a necessidade de retreinamento frequente, característica que a torna particularmente eficaz em ambientes dinâmicos.

Nesse cenário, o estudo de Lu *et al*. (2024) propõe o desenvolvimento de um assistente inteligente baseado em LLMs, funcionando como uma interface de linguagem natural para execução de tarefas complexas de simulação. A estrutura desenvolvida incorpora mecanismos robustos de tratamento de exceções, visando garantir a confiabilidade e a consistência lógica das simulações — aspecto fundamental na área de energia. Além disso, o trabalho reconhece os riscos da flexibilidade excessiva dos LLMs em contextos técnicos e propõe soluções para evitar o uso de parâmetros inválidos ou execuções incorretas. O sistema foi validado com diferentes LLMs e avaliado por meio de métricas de taxa de conclusão de tarefas, demonstrando avanços na interação inteligente e confiável com ferramentas técnicas específicas. Essa abordagem destaca-se por ir além da geração textual genérica, configurando-se como um modelo de *chatbot* especializado voltado à simulação e análise técnica em ambientes críticos.

Oliveira (2024) propõe o desenvolvimento de um assistente inteligente baseado em *chatbot* voltado para o setor elétrico, também utilizando um LLM aliado à técnica de indexação RAPTOR (*Recursive Abstractive Processing for Tree-Organized Retrieval*). O objetivo principal é facilitar o acesso e a compreensão das normas técnicas por profissionais da área elétrica, desde a elaboração de projetos até a manutenção e inspeção de instalações.

A metodologia do trabalho de Oliveira (2024) incluiu a coleta e indexação de documentos normativos, conversão dos textos em *embeddings*, armazenamento em banco de dados vetorial e implementação de um *pipeline* de recuperação e geração de respostas. Os resultados demonstraram alto desempenho nas métricas de *Answer Relevancy* e *Faithfulness*, indicando que as respostas geradas são relevantes e precisas, mesmo com a presença de informações adicionais nos contextos recuperados — característica da técnica RAPTOR que pode impactar a métrica de

Context Relevancy. Ainda assim, o sistema mostrou-se eficaz no suporte técnico, promovendo maior eficiência, agilidade e conformidade regulatória no setor elétrico.

Os trabalhos analisados apresentam o uso de *chatbots* inteligentes aplicados a diferentes contextos, como educação, suporte técnico, simulação e normas do setor elétrico. Em comum, destacam-se o uso de LLMs, técnicas de recuperação de informações e a busca por respostas mais precisas e atualizadas, geralmente por meio de abordagens como RAG ou RAPTOR.

A aplicação desenvolvida neste trabalho se diferencia por integrar múltiplos recursos — como recuperação textual e de imagens, fluxo condicional e memória de contexto — em uma única arquitetura voltada para o setor de energia. Além disso, destaca-se pelo foco em desempenho, com análise detalhada do tempo de resposta, e pela interface multimodal, que enriquece a experiência do usuário em ambientes operacionais.

3. METODOLOGIA

O *chatbot* proposto foi desenvolvido com base em uma arquitetura que combina técnicas de *Embedding*, RAG e o uso de LLMs. Essa abordagem visa criar um sistema capaz de compreender perguntas formuladas em linguagem natural, recuperar as informações mais relevantes de uma base de dados textual e gerar respostas precisas e humanizadas.

Para estruturar o fluxo de operação do *chatbot*, foi utilizada a biblioteca *LangGraph*, que permite a construção de diagramas de estados para o controle do ciclo de vida da aplicação, desde a recepção da entrada do usuário até a geração da resposta final. A seguir, descrevem-se detalhadamente os principais componentes da aplicação desenvolvida.

3.1. Large Language Models (LLMs)

As LLMs são modelos de aprendizado profundo treinados com grandes volumes de dados textuais, com o objetivo de compreender e gerar linguagem natural de forma contextualizada e coerente. Esses modelos baseiam-se majoritariamente na arquitetura

transformer, sendo capazes de realizar uma ampla gama de tarefas em Processamento de Linguagem Natural (PLN), como resposta a perguntas, geração de texto, sumarização e tradução automática (MINAEE *et al*., 2024).

Neste trabalho, optou-se pela utilização do modelo *LLaMA 3.3-70B*, desenvolvido pela *Meta AI*. Essa versão é composta por 70 bilhões de parâmetros e foi projetada para fornecer respostas mais precisas, fluentes e alinhadas às instruções do usuário. Este modelo se destaca pela eficiência no uso de memória e desempenho competitivo em tarefas de inferência, além de suportar instruções em linguagem natural com alta fidelidade semântica (META AI, 2024). Sua robustez o torna uma escolha adequada para aplicações que exigem compreensão profunda do contexto e geração confiável de respostas.

A escolha deste modelo foi motivada por seu caráter *open-source*, que possibilita maior controle, personalização e integração ao sistema proposto. Além disso, o modelo demonstrou desempenho compatível com os requisitos de qualidade e confiabilidade esperados, apresentando uma boa relação entre custo computacional e precisão nas respostas. Tais características o tornaram uma escolha adequada neste contexto em relação a outros modelos disponíveis no momento.

Apesar de seu potencial, o uso de LLMs, como o *LLaMA 3.3-70B*, apresenta desafios importantes, como o risco de geração de informações imprecisas (*hallucination*) (MINAEE *et al*., 2024), a presença de vieses nos dados de treinamento e a alta demanda computacional (BENDER *et al*., 2021). Para minimizar esses riscos e aumentar a confiabilidade das respostas, adota-se a técnica de RAG, que será apresentada em detalhe nas seções seguintes.

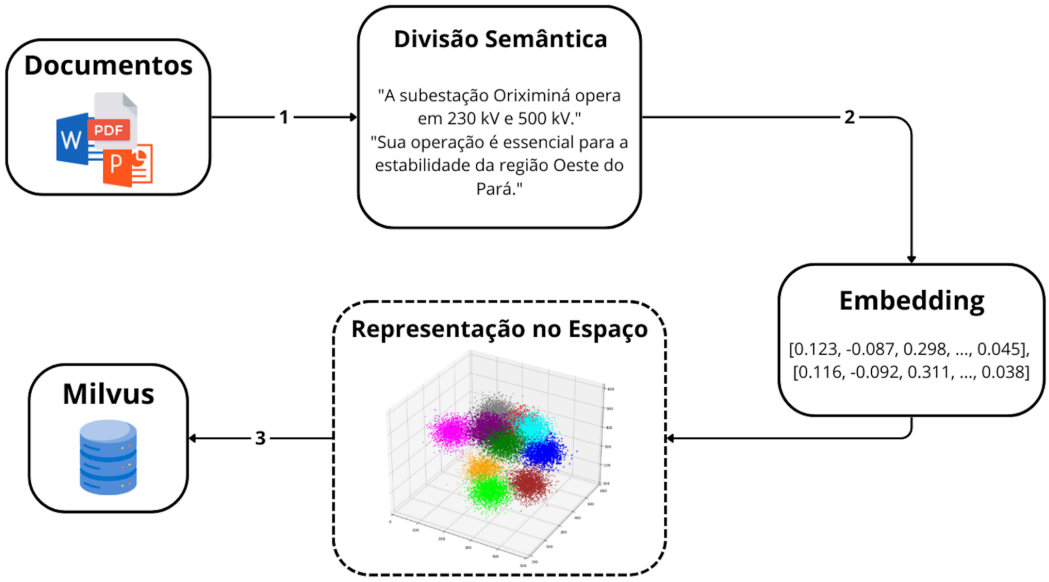
3.2. Embedding

A técnica de *embedding* consiste em transformar textos em vetores numéricos de alta dimensionalidade, preservando relações semânticas no espaço vetorial (RODRIGUES *et al*., 2020). Esse processo permite que palavras, sentenças ou documentos com significados semelhantes sejam representados por vetores próximos entre si, o que é fundamental para aplicações como busca semântica, sistemas de recomendação e classificação de textos (TANG, 2020).

A comparação entre esses vetores é geralmente realizada por meio da similaridade de cosseno, que avalia o ângulo entre dois vetores: quanto menor o ângulo, maior a similaridade entre os conteúdos representados (ZHOU *et al.*, 2020). Diferentemente de abordagens baseadas apenas em palavras-chave, os *embeddings* são capazes de capturar o significado contextual das frases, mesmo quando diferentes termos são utilizados, promovendo uma recuperação mais precisa e relevante das informações (NOVOTNÝ *et al.*, 2020).

O fluxograma na Figura 1 ilustra esse processo: documentos são lidos e divididos em trechos semânticos, que são então convertidos em vetores numéricos (*embeddings*). Esses vetores são armazenados em um banco vetorial, como o *Milvus*, permitindo buscas por similaridade semântica.

Figura 1 – Fluxograma do Processo de *Embedding*.



Fonte: Elaborado pelo próprio autor

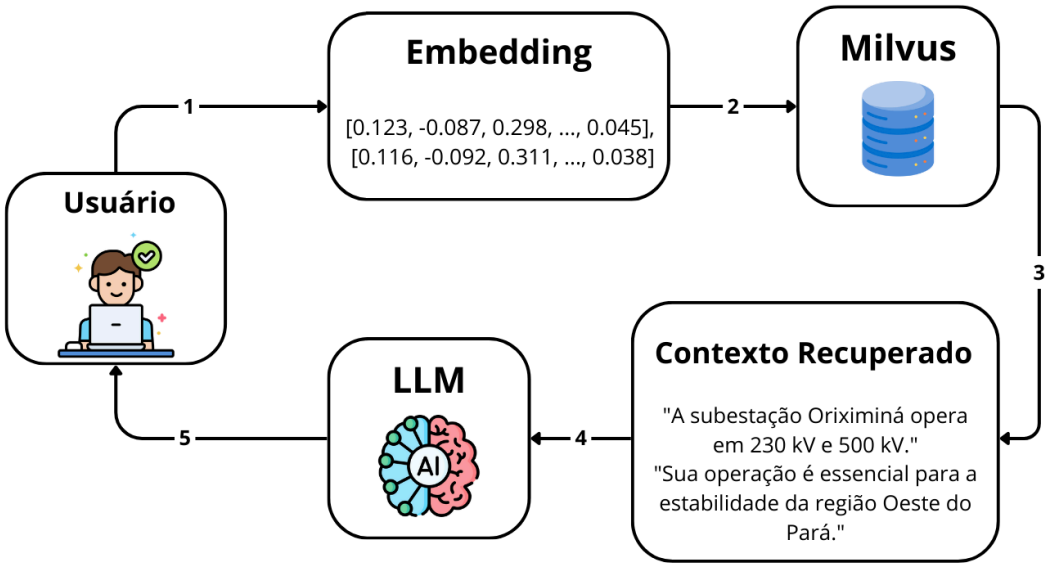
3.3. Recuperação Aumentada por Geração (RAG)

O sistema adota a abordagem RAG para melhorar a qualidade e a contextualização das respostas. Essa técnica combina a recuperação eficiente de trechos textuais relevantes com a capacidade de geração de linguagem natural de uma LLM (LEWIS *et al.*, 2020).

No fluxo RAG, a pergunta do usuário é primeiro convertida em *embedding* é utilizada para buscar segmentos de documentos semanticamente similares na base vetorial. Os trechos recuperados são então incorporados como contexto para a LLM, que gera uma resposta coerente e informada, fundamentada nos dados recuperados (HUANG; HUANG, 2024).

Essa combinação permite que o sistema supere limitações comuns a modelos geradores puros, que podem inventar informações ou responder de forma genérica, ao incorporar diretamente conhecimento extraído da base de dados indexada. O RAG, portanto, melhora a precisão, relevância e atualidade das respostas geradas (GAO *et al*., 2023). A Figura 2 ilustra esse fluxo de forma esquemática, apresentando as etapas que compõem a arquitetura RAG utilizada no sistema.

Figura 2 – Fluxograma de funcionamento do RAG.



Fonte: Elaborado pelo próprio autor

3.4. Geração da Base de Dados

A base de dados do sistema foi construída a partir da ingestão de documentos localizados em um diretório, utilizando o componente *DirectoryLoader* da biblioteca *LangChain*. Esse componente permite o carregamento de arquivos com diferentes formatos, padronizando o conteúdo para processamento posterior. A coleção utilizada era composta por 10 documentos técnicos contendo orientações detalhadas sobre os

procedimentos que devem ser executados pelos controladores em diferentes cenários operacionais, incluindo etapas de manobras, protocolos de segurança e instruções para situações de emergência.

Para viabilizar uma recuperação eficiente e contextualizada, os documentos foram segmentados em unidades semânticas menores, como parágrafos ou sentenças. Essa segmentação tem como objetivo estruturar a base em trechos com coesão temática suficiente para garantir relevância no momento da busca.

Cada segmento textual foi convertido em um vetor numérico de alta dimensionalidade, ou *embedding*, utilizando o modelo '*paraphrase-multilingual-mpnet-base-v2*' da biblioteca *SentenceTransformer*.

3.5. Base de Dados Vetorial

Os *embeddings* gerados foram armazenados em uma base vetorial construída com o *Milvus*, uma biblioteca especializada no gerenciamento e recuperação eficiente de dados vetoriais em larga escala. Esta possibilita buscas rápidas e escaláveis por similaridade semântica, sendo fundamental para a etapa de recuperação de informações do sistema (WANG *et al*., 2021).

3.6. Processamento e Indexação de Imagens

Além do texto, o sistema integra informações visuais para enriquecer as respostas. Cada imagem da base foi associada a um rótulo textual descritivo, organizado em formato JSON contendo o texto do rótulo e o diretório do arquivo correspondente. Esses rótulos também foram convertidos em *embeddings* utilizando o mesmo modelo semântico '*paraphrase-multilingual-mpnet-base-v2*', garantindo compatibilidade no espaço vetorial com os *embeddings* textuais.

Os vetores dos rótulos foram indexados em uma segunda tabela na base *Milvus*, dedicada exclusivamente à recuperação por similaridade semântica das imagens. Durante a inferência, essa base é consultada para sugerir imagens relevantes que complementam o conteúdo da resposta textual, promovendo uma experiência multimodal ao usuário.

3.7. Arquitetura Lógica e Fluxo de Execução com *LangGraph*

Para coordenar o fluxo de execução do *chatbot*, foi utilizada a biblioteca *LangGraph*, que permite modelar processos computacionais por meio de grafos direcionados. Cada nó representa uma operação funcional — como a recepção da pergunta, busca de contexto ou geração da resposta — enquanto as arestas definem transições condicionais entre essas etapas (LANGCHAIN Inc., 2025).

Esse modelo gráfico oferece flexibilidade para definir fluxos condicionais, ciclos e ramificações, essencial em sistemas de linguagem natural onde diferentes tipos de perguntas e dados disponíveis demandam tratamentos diferenciados. O fluxo é organizado nos seguintes nós principais:

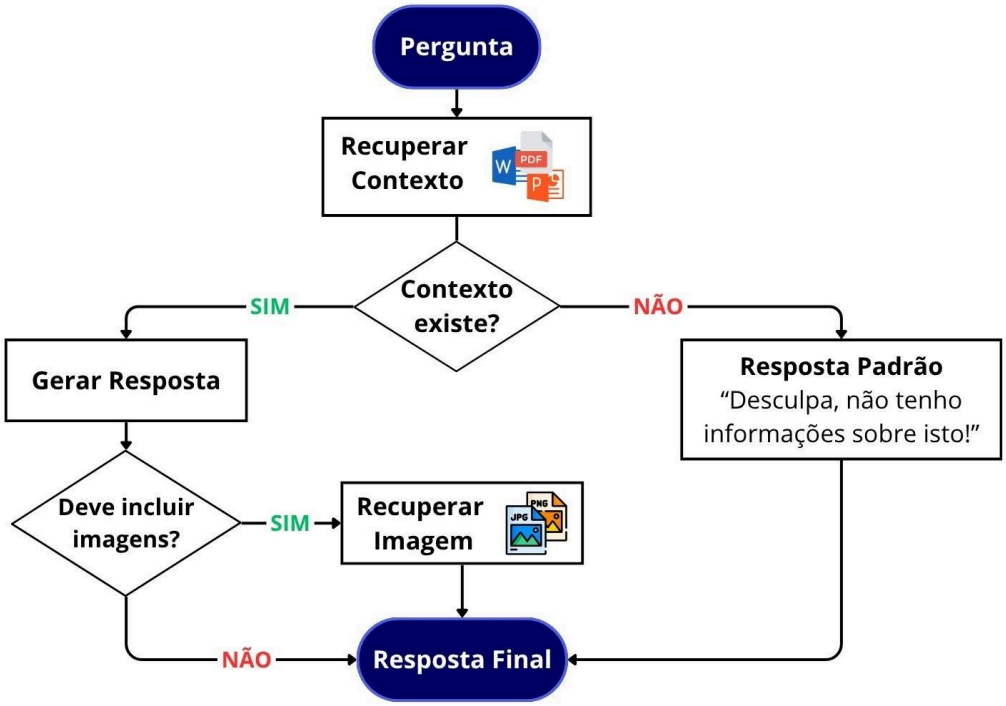
- **Recuperar Contexto:** recebe a pergunta do usuário, converte-a em *embedding* com o modelo *'paraphrase-multilingual-mpnet-base-v2'* e realiza busca por similaridade na base textual. Os segmentos mais relevantes são armazenados como contexto para geração da resposta.
- **Gerar Resposta:** utiliza a pergunta original e o contexto recuperado para alimentar uma LLM que, guiada por *prompts* pré-definidos, gera uma resposta textual coerente, contextualizada e natural.
- **Recuperar Imagens:** recebe a resposta gerada pela LLM, converte-a em *embedding* semântico e consulta a base vetorial de rótulos de imagens para identificar e retornar a imagem mais relevante para enriquecer a resposta.

Além disso, blocos condicionais no grafo avaliam a necessidade de incluir imagens na resposta (verificando presença de palavras-chave na pergunta) e a relevância do contexto recuperado para decidir entre gerar uma resposta elaborada ou apresentar uma mensagem padrão caso não haja dados suficientes.

Outro diferencial da biblioteca utilizada é seu mecanismo interno de memória de estado, capaz de registrar os estados de todos os nós do grafo a cada iteração. Esse recurso permite a construção de um histórico de conversas, possibilitando que o *chatbot* tenha acesso às interações anteriores para formular respostas mais

contextualizadas em novos diálogos (LANGCHAIN Inc., 2025). Como resultado, a experiência do usuário é aprimorada, uma vez que o sistema pode manter a coerência e a continuidade durante sessões de conversa prolongadas. A Figura 3 ilustra o fluxo construído para o chat.

Figura 3 – Fluxograma de funcionamento do Chat.



Fonte: Elaborado pelo próprio autor

3.8. Interface Gráfica

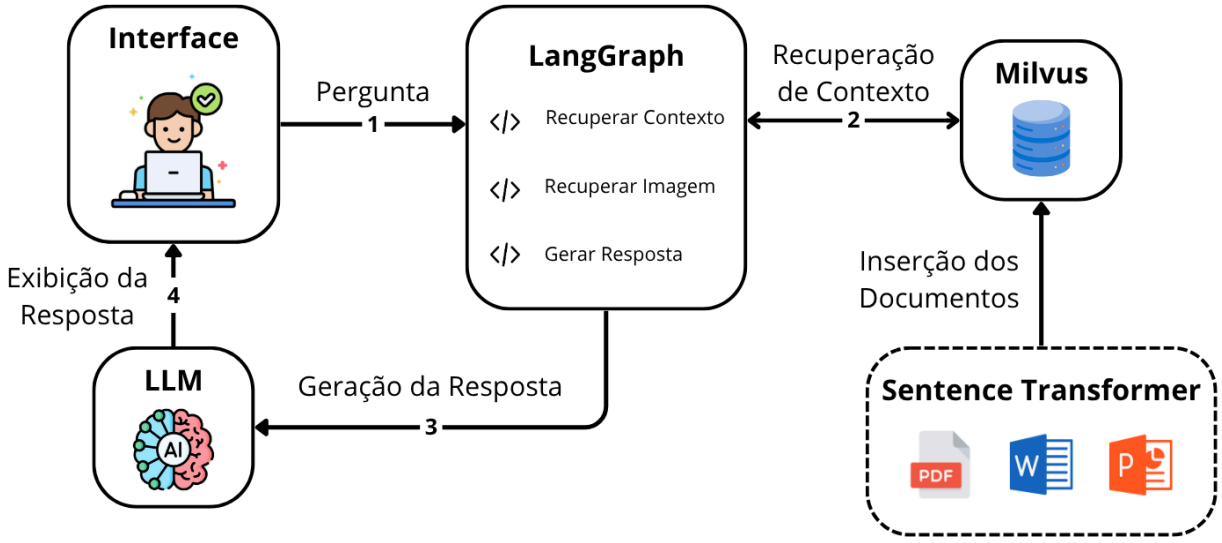
Com o objetivo de aprimorar a experiência do usuário e assegurar uma apresentação interativa, responsiva e de fácil utilização, o *front-end* do sistema de chat foi desenvolvido utilizando HTML, CSS e JavaScript, enquanto o *back-end* foi implementado com o *framework Flask*. Essa abordagem permitiu a criação de uma aplicação *web* com maior controle sobre a interface, garantindo flexibilidade na personalização e integração dos componentes visuais com a lógica do sistema.

Além disso, o *chat* conta com a possibilidade de criar múltiplas conversas, e o histórico é garantido pelo armazenamento no cache do navegador, permitindo que o usuário retome interações anteriores de forma rápida e contínua.

3.9. Estrutura Geral do ChatBot

Na Figura 4 é possível visualizar a função de cada técnica e biblioteca empregada na criação do *chatbot*. A arquitetura é composta por quatro etapas principais. Inicialmente, o usuário interage com a aplicação por meio da interface desenvolvida, onde insere sua pergunta. Essa solicitação é processada pelo módulo *LangGraph*, responsável por coordenar o fluxo de informações e realizar a recuperação de contexto. Para isso, o *LangGraph* consulta a base vetorial *Milvus*, que armazena representações semânticas de documentos previamente processados. A vetorização desses documentos (PDF, *Word*, *PowerPoint*, entre outros) é realizada por meio do modelo *Sentence Transformer*, possibilitando buscas semânticas eficientes. Com o contexto recuperado, o *LangGraph* encaminha as informações relevantes a um LLM, que é responsável pela geração da resposta textual. Por fim, essa resposta é enviada de volta à interface, onde é exibida ao usuário de forma clara e objetiva.

Figura 4 – Estrutura do Chat.



Fonte: Elaborado pelo próprio autor

4. RESULTADOS E DISCUSSÕES

A implementação do *chatbot* resultou em uma interface interativa e funcional, voltada para facilitar o acesso às informações operacionais de uma empresa do setor de energia. A Figura 5 ilustra a interface final do sistema, permitindo que o usuário

realize consultas em linguagem natural, recebendo respostas precisas e contextualizadas com base nos dados indexados.

Figura 5 – Interface final do Chat.

Fonte: Elaborado pelo próprio autor

Para avaliar a precisão e a capacidade de compreensão do sistema, foi realizada uma consulta sobre as conexões da Subestação Oriximiná. O chatbot respondeu listando corretamente todas as linhas de transmissão (LTs) associadas à subestação, com suas respectivas tensões e destinos, além de fornecer uma explicação complementar sobre as interligações com as subestações Juruti, Jurupari e Silves.

Ao comparar a resposta gerada com o conteúdo original, observa-se uma correspondência exata com os dados estruturados: todas as seis LTs foram identificadas corretamente, os níveis de tensão foram mencionados com precisão. Além disso, o sistema devolveu ao usuário um diagrama da rede base que, embora não estivesse presente nos dados originais, foi incluído como elemento visual de apoio. Este recurso multimodal oferece ao usuário uma representação espacial da rede de transmissão na região do Pará.

Esse resultado evidencia a capacidade do *chatbot* em interpretar a intenção do usuário, recuperar e organizar informações técnicas com precisão, e apresentar a

resposta de forma clara e visualmente compreensível. A integração entre a linguagem natural e o componente gráfico reforça a eficácia da abordagem adotada, sobretudo em contextos operacionais que exigem interpretação rápida e precisa de dados complexos.

Os testes também demonstraram que a integração da técnica RAG com o modelo escolhido atuou como um mecanismo eficaz para mitigar a geração de respostas imprecisas. Ao fornecer à LLM trechos relevantes previamente extraídos da base vetorial, o sistema reduziu significativamente a ocorrência de *hallucinations* e aumentou a confiabilidade das respostas. Essa integração direta entre recuperação e geração mostrou-se a principal estratégia para assegurar precisão técnica nas respostas.

Além disso, foram realizados testes para medir o tempo médio de resposta do chatbot. Para isso, foram feitas trinta perguntas representativas da base de informações do sistema, a qual, durante os testes, era composta por 10 documentos técnicos contendo orientações detalhadas sobre os procedimentos que devem ser executados pelos controladores em diferentes cenários operacionais, incluindo etapas de manobras, protocolos de segurança e instruções para situações de emergência. O objetivo foi avaliar a performance em diferentes etapas do processamento das solicitações. A Tabela 1 apresenta os tempos médios obtidos para cada etapa principal do fluxo de resposta.

Tabela 1 – Tempos médios por nó após 30 iterações.

Etapa	**Tempo Médio (s)**
Recuperar Contexto	0.1528
Condicional Contexto	0.0000
Gerar Resposta	1.1685
Condicional Imagem	0.0001
Recuperar Imagens	0.3043
Tempo Médio Total	**1.6257**

Fonte: Elaborado pelo próprio autor

Observa-se que a maior parte do tempo de processamento está concentrada na etapa de geração da resposta, que corresponde em média a cerca de 1,17 segundos por interação. A recuperação de imagens também impacta o tempo total, com cerca de 0,30 segundos. As etapas condicionais, por sua vez, têm tempos desprezíveis, indicando que a lógica condicional não adiciona atrasos significativos.

No geral, o tempo médio total estimado para o sistema responder a uma pergunta é de aproximadamente 1,63 segundos, um valor considerado satisfatório para garantir uma experiência fluida e interativa para o usuário. Esse desempenho indica que o chatbot é capaz de processar as informações e gerar respostas em tempo real, minimizando a sensação de espera.

5. SEGURANÇA E PRIVACIDADE

A segurança da informação, a privacidade dos dados e a conformidade regulatória constituem elementos fundamentais no projeto do chatbot proposto. O tratamento das informações é orientado pelos princípios da Lei Geral de Proteção de Dados Pessoais (LGPD), garantindo bases legais adequadas, limitação da finalidade e minimização dos dados utilizados ao estritamente necessário para o funcionamento do sistema (BRASIL, 2018).

O acesso ao chatbot é estruturado a partir de mecanismos de autenticação e autorização baseados em perfis de usuário, assegurando a segregação de funções e a restrição de acesso às informações conforme as atribuições de cada grupo. Essa organização está alinhada às práticas de gestão da segurança da informação estabelecidas pela norma ISO/IEC 27001 (ORGANIZAÇÃO INTERNACIONAL DE NORMALIZAÇÃO, 2022).

6. CONCLUSÕES

Tendo isso em vista, o *chatbot* desenvolvido se consolida não apenas como um assistente de consulta, mas como uma ferramenta estratégica para a gestão da informação em tempo real, contribuindo para uma tomada de decisão mais ágil, precisa e embasada, especialmente em cenários críticos de operação no setor elétrico.

A fluidez do fluxo de interação, viabilizada pela biblioteca *LangGraph*, proporciona uma experiência intuitiva, com respostas rápidas, claras e adaptadas à linguagem técnica dos usuários. Essa característica representa um avanço significativo em relação aos métodos tradicionais de consulta a documentos, muitas vezes dispersos em sistemas internos ou arquivos manuais.

Do ponto de vista de desempenho, o sistema demonstrou robustez na recuperação semântica, mesmo diante de perguntas formuladas com termos distintos dos presentes na base de dados. Esse resultado é atribuído à integração eficaz das técnicas de *embedding* e RAG, que garantem respostas mais relevantes, contextualizadas e informativas.

Assim, conclui-se que a arquitetura proposta representa um avanço tecnológico aplicável a contextos industriais, com potencial de expansão para outras áreas mediante a incorporação de novas fontes de dados e funcionalidades, como aquelas previstas nos trabalhos futuros.

7. TRABALHOS FUTUROS

Como trabalho futuro, propõe-se a implementação do *chatbot* em ambiente local, por meio da plataforma *Ollama*, com o objetivo de garantir maior controle sobre a execução do modelo, aumentar a segurança dos dados e permitir o uso de informações sensíveis sem risco de vazamentos.

Além disso, busca-se ampliar o acesso a dados operacionais por meio da integração do chatbot com o banco de dados corporativo. Para isso, será necessário desenvolver uma camada de integração entre a LLM e o banco de dados relacional, permitindo que comandos em linguagem natural sejam automaticamente convertidos em consultas SQL, viabilizando a recuperação dinâmica e precisa de informações diretamente do sistema de dados da empresa.

Em complemento a essas melhorias, estão previstas evoluções relacionadas à avaliação do desempenho e da confiabilidade do sistema. Serão incorporadas métricas objetivas, como *faithfulness*, *answer relevancy* e *context relevancy*, possibilitando verificar a aderência das respostas aos trechos recuperados e a adequação do contexto utilizado. Adicionalmente, métricas clássicas de avaliação, como *precision*,

recall e *F1-score*, serão aplicadas a um conjunto de consultas previamente validadas por especialistas. Também será analisada a taxa de alucinação e a ocorrência de erros críticos, especialmente aqueles associados à geração de instruções operacionais incorretas.

Outra ampliação prevista refere-se à expansão da base de conhecimento do chatbot, incluindo documentos normativos, procedimentos operacionais, registros de ocorrência e ícones ou diagramas padronizados, associados a mecanismos de controle de versões. Essa abordagem visa aumentar a cobertura informacional, a rastreabilidade das fontes e a confiabilidade das respostas fornecidas.

No aspecto arquitetural, serão implementadas e avaliadas diferentes estratégias de recuperação e geração de respostas, contemplando um *baseline* baseado em BM25 combinado com RAG simples, bem como abordagens mais avançadas, como *Graph-RAG* e *RAPTOR*, além da aplicação de técnicas de *re-ranking*. A comparação entre essas arquiteturas permitirá analisar ganhos de desempenho, qualidade das respostas e redução de alucinações.

Por fim, os aspectos de usabilidade e experiência do usuário serão aprofundados por meio de testes com usuários-alvo, como controladores e operadores. Esses testes incluirão a aplicação de um checklist de segurança operacional, com foco na prevenção de instruções potencialmente perigosas, bem como a coleta de feedback estruturado, possibilitando o aprimoramento contínuo do sistema.

8. AGRADECIMENTOS

Os autores agradecem ao Programa de Educação Tutorial de Engenharia Elétrica (PET-EE) da Universidade Federal do Pará (UFPA) e à concessionária Equatorial Energia Pará.

9. DECLARAÇÃO DO USO DE INTELIGÊNCIA ARTIFICIAL

Durante o desenvolvimento deste trabalho, foi utilizada a (IA) para auxiliar em algumas etapas da pesquisa e elaboração deste documento, conforme segue:

- **Ferramenta de IA utilizada**: Chat GPT
    - **Objetivo do uso**: A IA foi empregada para auxiliar na elaboração de resumos, sugestões de tópicos, revisão gramatical, organização de referências bibliográficas e aprimoramento da clareza textual.

- **Ferramenta de IA utilizada**: Manus
    - **Objetivo do uso**: A inteligência artificial foi empregada para identificar e corrigir bugs, além de otimizar trechos de código para melhorar desempenho e eficiência.

A autoria do conteúdo original e as análises realizadas neste trabalho são de responsabilidade exclusiva do autor. A utilização da IA não substitui a análise crítica, interpretação ou conclusões da pesquisa. A IA foi empregada como ferramenta auxiliar para facilitar algumas etapas do processo de escrita e análise. Declaro, portanto, que o uso da IA foi realizado dentro das normas éticas e acadêmicas, com o intuito de melhorar a qualidade e precisão do trabalho, sem comprometer a integridade e a originalidade do conteúdo.

André Oliveira Carvalho da Silva - Autor do Trabalho
Belém- PA, 01 de dezembro de 2025.

REFERÊNCIAS

ALBERTIN, Alberto Luiz; ALBERTIN, Rosa Maria de Moura. Transformação digital: gerando valor para o “novo futuro”. *GV Executivo*, São Paulo, v. 20, n. 1, p. 26–29, 2021.

ANTÔNIO, Marcos José Pereira de; ALVES, João Paulo; NASCIMENTO, Diego Torres do; GARRIDO, Maria Cristina de Melo. Automação logística, inteligência artificial e Indústria 4.0: considerações éticas em um cenário de inovação tecnológica. *Mobicities – Journal of Urban Mobility, Logistics and Sustainable Smart Cities*, v. 2, n. 1, 2025.

BENDER, Emily M. Sobre os perigos dos “papagaios estocásticos”: os modelos de linguagem podem ser grandes demais? In: *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FAccT 2021)*. Nova York: ACM, 2021. p. 610–623.

BRASIL. Lei nº 13.709, de 14 de agosto de 2018. Lei Geral de Proteção de Dados Pessoais (LGPD). Diário Oficial da União, Brasília, DF, 15 ago. 2018. Disponível em: https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm. Acesso em: 21 maio 2025.

COSTA NETO, Luiz Gonzaga da; CAMPOS, Fernando César de. Oportunidades de aplicação de business intelligence no contexto da Indústria 4.0: revisão sistemática da literatura (2015–2020). *Exacta*, São Paulo, v. 21, n. 2, p. 503–519, 2023.

DELBIANCO, Natália Ribeiro; VALENTIM, Marta Lígia Pomim. Sociedade da informação e mídias sociais no contexto da comunicação científica. *AtoZ: Novas Práticas em Informação e Conhecimento*, Curitiba, v. 11, p. 1–11, 2022.

FREIRE, André Augusto de Almeida. IFPBBot: um chatbot acadêmico. 2024. Trabalho de Conclusão de Curso (Graduação) – Instituto Federal da Paraíba, Paraíba, 2024.

GAÄL, Luís Paulo Martins; MARTINS, Matheus Silva. Acesso aberto no contexto da pesquisa em ciência da informação. *Transinformação*, Campinas, v. 34, e220016, 2022.

GAO, Yifan et al. Geração aumentada por recuperação para modelos de linguagem de grande escala: uma revisão. *arXiv*, 2023. Disponível em: https://arxiv.org/abs/2312.10997. Acesso em: 21 maio 2025.

HERNANDEZ, Edson Fernando Teixeira; TOLEDO, Natália Karla Almeida de. Crimes cibernéticos: efeitos revolucionários diante de uma legislação em constante evolução. *Revista Jurídica da UniFil*, Londrina, v. 17, n. 17, p. 72–84, 2021.

HUANG, Yifan; HUANG, Jing. Uma revisão sobre geração de texto aumentada por recuperação aplicada a grandes modelos de linguagem. *arXiv*, 2024. Disponível em: https://arxiv.org/abs/2404.10981. Acesso em: 21 maio 2025.

INTERNATIONAL ORGANIZATION FOR STANDARDIZATION. ISO/IEC 27001:2022 — Tecnologia da informação — Técnicas de segurança — Sistemas de gestão da segurança da informação — Requisitos. Genebra: ISO, 2022. Disponível em: https://www.iso.org/standard/82875.html. Acesso em: 21 maio 2025.

LANGCHAIN INC. LangGraph: framework open-source para o desenvolvimento de agentes de inteligência artificial adaptáveis e confiáveis. 2025. Disponível em: https://www.langchain.com/langgraph. Acesso em: 21 maio 2025.

LARUSSA FILHO, C. G.; BORGES, J. H. G.; PORTA, C. H. M.; FLORIAN, F. Desenvolvimento de chatbot para consulta de informações acadêmicas do aluno. *RECIMA21 – Revista Científica Multidisciplinar*, v. 3, n. 12, e3122364, 2022.

LEE, Hyeon-Cheol. Desenvolvimento de um chatbot baseado em RAG e LLM para aprimoramento do suporte técnico. 2024.

LEWIS, Patrick et al. Geração aumentada por recuperação para tarefas de processamento de linguagem natural intensivas em conhecimento. *arXiv*, 2020. Disponível em: https://arxiv.org/abs/2005.11401. Acesso em: 21 maio 2025.

LUGLI, Victor Augusto; LUCCA FILHO, João de. O uso de chatbots para a excelência no atendimento. *Revista Interface Tecnológica*, Taquaritinga, v. 17, n. 1, p. 205–218, 2020.

LU, Lei et al. Assistente inteligente para simulação de sistemas de potência: um chatbot integrando modelos de linguagem e modelos de domínio. 2024.

MARTINS, André; PAIVA, Paulo Augusto. Inteligência artificial nas empresas: estudo de caso do chatbot LZ. *Revista Brasileira de Gestão e Inovação*, Caxias do Sul, v. 12, n. 1, 2025.

MENEGUELLO, Amanda Silva. Chatbot educacional: desenvolvimento de um chatbot para estudantes e educadores. 2023.

META AI. Introdução ao LLaMA 3: modelos fundamentais abertos e ajustados por instrução. 2024. Disponível em: https://ai.meta.com/llama/. Acesso em: 21 maio 2025.

MINAEE, Shervin et al. Modelos de linguagem de grande escala: uma revisão. *arXiv*, 2024. Disponível em: https://arxiv.org/abs/2402.06196. Acesso em: 21 maio 2025.

MORAIS, Márcio de Oliveira. A evolução da qualidade na Indústria 4.0. *Research, Society and Development*, v. 9, n. 10, p. e3929108634, 2020.

NOVOTNÝ, David. Classificação de textos com regularização por embeddings e medida de similaridade suave. *arXiv*, 2020. Disponível em: https://arxiv.org/abs/2003.05019. Acesso em: 21 maio 2025.

OLIVEIRA, Natan Silva de. Desenvolvimento de um assistente chatbot inteligente para instalações elétricas baseado em modelo de linguagem de grande escala (LLM). 2024.

OLIVEIRA, Victor Ferreira de. Arquitetura para integração e análise de dados baseada no padrão Indústria 4.0. 2023.

PALLIN, Guilherme et al. Vacabot: um chatbot que emprega linguagem natural no ambiente de viagens. *Anais do Congresso Brasileiro de Iniciação Científica*, v. 1, n. 2, p. 77–82, 2024.

PARKAR, Rohan et al. Chatbot interativo acadêmico baseado em inteligência artificial e web. 2021.

PERFETTO, Flávia Vieira; REIS, Sérgio Gonçalves de Oliveira; PALETTA, Francisco Carlos. Gestão da informação digital: caminhos possíveis. *RDBCI*, Campinas, v. 21, e023005, 2023.

RODRIGUES, Tiago et al. Modelos de linguagem e embeddings em português: avaliação em tarefas de similaridade semântica. In: *Experimental IR Meets Multilinguality, Multimodality, and Interaction*. Cham: Springer, 2020. p. 295–307.

TANG, Li. Combinação do BERT com embeddings de sentido do WordNet para previsão de mudanças graduais de similaridade lexical. In: *Proceedings of the Fourteenth Workshop on Semantic Evaluation*. Barcelona: ACL, 2020. p. 166–170.

WANG, Jianguo et al. Milvus: um sistema de gerenciamento de dados vetoriais desenvolvido para grandes volumes. In: *Proceedings of the International Conference on Management of Data (SIGMOD 2021)*. Nova York: ACM, 2021.

ZHOU, Jie et al. Identificação estrutural bioinspirada em embeddings de linguagem. *arXiv*, 2020. Disponível em: https://arxiv.org/abs/2009.02459. Acesso em: 21 maio 2025.

ANEXO A - ARTIGO COMPLETO PUBLICADO NO INDUSCON 2025

Chatbot Inteligente para Auxílio à Tomada de Decisão em Empresa do Setor de Energia

André Oliveira Carvalho da Silva
Instituto de Tecnologia
Universidade Federal do Pará
Belém, Brasil
andre.carvalho.silva@itec.ufpa.br

Emanuele Duarte Silva
Instituto de Tecnologia
Universidade Federal do Pará
Belém, Brasil
emanuele.silva@itec.ufpa.br

Elen Priscila de Souza Lobato
Instituto de Tecnologia
Universidade Federal do Pará
Belém, Brasil
elen.lobato@itec.ufpa.br

Wellington da Silva Fonseca
Instituto de Tecnologia
Universidade Federal do Pará
Belém, Brasil
fonseca@ufpa.br

Lusiane Pereira Fonseca
Gerência de Operações
Equatorial Energia
Belém, Brasil
lusiane.fonseca@equatorialenergia.com.br

***Resumo*—Este trabalho apresenta o desenvolvimento de um chatbot inteligente voltado para auxiliar a tomada de decisão operacional em uma concessionária de energia elétrica da região Norte do Brasil. A aplicação utiliza Modelos de Linguagem de Larga Escala (LLMs), *embeddings* semânticos e a técnica de Recuperação Aumentada por Geração (RAG) para oferecer respostas precisas, contextualizadas e multimodais, a partir de uma base de dados textual e visual indexada. A arquitetura do sistema foi estruturada com o uso da biblioteca *LangGraph*, que permite o controle dinâmico do fluxo conversacional, e do *Milvus*, um sistema de gerenciamento de dados vetoriais. A interface desenvolvida com HTML, CSS e JavaScript proporciona uma experiência interativa e acessível ao usuário. Testes demonstraram precisão nas respostas e tempo médio de resposta inferior a dois segundos, evidenciando o potencial do chatbot como ferramenta estratégica para aumento da eficiência operacional no setor elétrico. O estudo destaca a viabilidade técnica e o impacto positivo da aplicação de IA generativa em ambientes industriais críticos.**

***Palavras-chave*—Chatbot, Setor Elétrico, Recuperação Aumentada por Geração, Modelos de Linguagem de Larga Escala, Tomada de Decisão Operacional,**

I. Introdução

O acesso à informação desempenha um papel central na sociedade, evoluindo desde a leitura de jornais impressos até a utilização de ferramentas baseadas em Inteligência Artificial (IA) [1]. Esse processo de transformação reflete nos avanços tecnológicos e nas mudanças nos modos de produção, circulação e consumo da informação. No passado, o acesso era limitado por fatores como tempo, espaço e meios físicos de distribuição [2]. Com o advento da internet e, posteriormente, das tecnologias digitais avançadas, como o armazenamento em nuvem, os mecanismos de busca e os sistemas inteligentes, a informação tornou-se mais acessível, dinâmica e descentralizada [3].

Atualmente, vivencia-se uma era marcada pela velocidade com que os dados são gerados e compartilhados, exigindo novas habilidades cognitivas e digitais por parte dos indivíduos e das organizações [4]. O conhecimento, que antes era acumulado em bibliotecas, arquivos físicos e instituições formais, passou a circular por redes interconectadas e plataformas digitais que operam em tempo real [5]. Nesse novo cenário digital, a informação deixa de ser apenas um suporte para decisões estratégicas e passa a ocupar uma posição central como recurso de alto valor para a inovação, a competitividade e o desenvolvimento social.

Em meio a essa transformação, ferramentas baseadas em Inteligência Artificial (IA), como sistemas de recomendação, algoritmos preditivos e assistentes virtuais, tornam-se essenciais para lidar com o volume e a complexidade dos dados contemporâneos [6]. Essas tecnologias facilitam o acesso à informação, otimizando sua filtragem, análise e personalização e promovem uma interação mais eficiente e alinhada às necessidades específicas dos usuários. Essa capacidade de adaptação, orientada por dados, exemplifica uma das principais características da chamada Indústria 4.0 [7].

A Indústria 4.0 é marcada pela integração de tecnologias digitais aos processos produtivos, possibilitando automação inteligente, conectividade em tempo real e análise massiva de dados [8]. Nesse contexto, o uso de sistemas inteligentes favorece a tomada de decisões mais ágil, precisa e contextualizada, elevando os níveis de eficiência e competitividade organizacional [9]. As organizações, diante desse panorama, deixam de operar com sistemas exclusivamente executores e passam a adotar soluções tecnológicas capazes de interagir, interpretar e responder de forma mais humanizada [10].

Entre essas tecnologias, os $chatbots$ têm ganhado destaque por sua capacidade de intermediar a interação entre usuários e sistemas computacionais, utilizando linguagem natural [11]. Os $chatbots$ são programas baseados em regras pré-definidas e aprendizado de máquina, que permitem o acesso automatizado a informações armazenadas em nuvem ou em documentos corporativos [12]. Sua implementação promove ganhos significativos na gestão da informação, reduzindo a sobrecarga de atendimentos manuais, otimizando o tempo de resposta e ampliando o acesso a dados estratégicos de forma segura e eficiente.

As aplicações dos $chatbots$ são inúmeras e vêm se expandindo progressivamente em diversos setores, com

destaque para o ambiente empresarial [13]. No contexto corporativo, essas ferramentas têm sido utilizadas tanto no relacionamento com o cliente quanto no suporte interno às equipes e processos organizacionais [14]. Por meio da integração com repositórios digitais, sistemas internos e plataformas em nuvem, os *chatbots* permitem que colaboradores acessem conteúdos institucionais, normativas internas, procedimentos operacionais, históricos de atendimento, relatórios técnicos, entre outros documentos relevantes, utilizando linguagem natural e comandos simples [15].Essa funcionalidade reduz significativamente o tempo despendido na busca manual por informações, ao mesmo tempo em que minimiza erros e aumenta a eficiência operacional.

Dentro deste contexto, este estudo tem como objetivo apresentar o desenvolvimento e a aplicação de um *chatbot* voltado para uma concessionária de distribuição de energia da região Norte. O foco principal é proporcionar aos técnicos controladores um acesso mais rápido, direto e eficiente a informações operacionais relevantes, permitindo que consultas sejam feitas de forma ágil e concisa. Dessa maneira, busca-se otimizar os processos de tomada de decisão, especialmente em situações que demandam respostas imediatas e precisão no manejo dos dados técnicos e administrativos.

Para estruturar o desenvolvimento do *chatbot*, este trabalho foi organizado da seguinte forma: a Seção 2 apresenta os trabalhos relacionados, descrevendo as principais abordagens existentes. A Seção 3 descreve os conceitos fundamentais e as bibliotecas utilizadas no projeto. Em seguida, a Seção 4 discute os resultados obtidos, incluindo a interface desenvolvida e o desempenho do sistema. Por fim, a Seção 5 traz as considerações finais, sintetizando os principais achados e contribuições deste trabalho.

II. Desenvolvimento de ChatBot

Estudos presentes na literatura destacam o desenvolvimento de *chatbots* integrados a outras interfaces tecnológicas, com o intuito de facilitar o acesso à informação. Um exemplo disso é o projeto descrito por Parkar *et al.* [16], no qual foi desenvolvido um *chatbot* baseado na *web*, utilizando bibliotecas de Processamento de Linguagem Natural (PLN) e Aprendizado de Máquina para interagir com os usuários de forma automatizada. Considerando a constante atualização do currículo acadêmico, os autores implementaram um sistema que permite a modificação dinâmica do banco de dados por meio de uma interface desenvolvida em HTML (*HyperText Markup Language*) e PHP (*Hypertext Preprocessor*), possibilitando a atualização periódica das informações. A proposta do sistema visa melhorar a eficiência e a autonomia dos estudantes universitários, fornecendo dados essenciais diretamente da instituição de ensino.

Em estudo de Lee *et al.* [17] desenvolvido em parceria com a empresa canadense *OptiMicro Technologies Inc.*, foi apresentado um *chatbots* baseado em *Large Language Models* (LLMs) com capacidade de *Retrieval-Augmented Generation* (RAG), voltado para aprimorar o serviço de suporte técnico. O sistema foi construído utilizando a plataforma *Flowise AI* e testado com dados reais da empresa. A principal vantagem da abordagem RAG é sua habilidade de acessar dinamicamente bases de conhecimento externas, o que permite fornecer respostas atualizadas e sensíveis ao contexto, reduzindo o risco de alucinações típicas em modelos LLM tradicionais. Além disso, a RAG dispensa a necessidade de retreinamento frequente, característica que a torna particularmente eficaz em ambientes dinâmicos.

Nesse cenário, o estudo de Lu *et al.* [18] propõe o desenvolvimento de um assistente inteligente baseado em LLMs, funcionando como uma interface de linguagem natural para execução de tarefas complexas de simulação. A estrutura desenvolvida incorpora mecanismos robustos de tratamento de exceções, visando garantir a confiabilidade e a consistência lógica das simulações — aspecto fundamental na área de energia. Além disso, o trabalho reconhece os riscos da flexibilidade excessiva dos LLMs em contextos técnicos e propõe soluções para evitar o uso de parâmetros inválidos ou execuções incorretas. O sistema foi validado com diferentes LLMs e avaliado por meio de métricas de taxa de conclusão de tarefas, demonstrando avanços na interação inteligente e confiável com ferramentas técnicas específicas. Essa abordagem destaca-se por ir além da geração textual genérica, configurando-se como um modelo de chatbot especializado voltado à simulação e análise técnica em ambientes críticos.

Oliveira [19] propõe o desenvolvimento de um assistente inteligente baseado em *chatbot* voltado para o setor elétrico, também utilizando um LLM aliado à técnica de indexação RAPTOR (*Recursive Abstractive Processing for Tree-Organized Retrieval*). O objetivo principal é facilitar o acesso e a compreensão das normas técnicas por profissionais da área elétrica, desde a elaboração de projetos até a manutenção e inspeção de instalações.

A metodologia do trabalho de Oliveira [19] incluiu a coleta e indexação de documentos normativos, conversão dos textos em *embeddings*, armazenamento em banco de dados vetorial e implementação de um *pipeline* de recuperação e geração de respostas. Os resultados demonstraram alto desempenho nas métricas de *Answer Relevancy* e *Faithfulness*, indicando que as respostas geradas são relevantes e precisas, mesmo com a presença de informações adicionais nos contextos recuperados — característica da técnica RAPTOR que pode impactar a métrica de *Context Relevancy*. Ainda assim, o sistema mostrou-se eficaz no suporte técnico, promovendo maior eficiência, agilidade e conformidade regulatória no setor elétrico.

Os trabalhos analisados apresentam o uso de *chatbots* inteligentes aplicados a diferentes contextos, como educação, suporte técnico, simulação e normas do setor elétrico. Em comum, destacam-se o uso de LLMs, técnicas de recuperação de informações e a busca por respostas mais precisas e atualizadas, geralmente por meio de abordagens como RAG ou RAPTOR.

A aplicação desenvolvida neste trabalho se diferencia por integrar múltiplos recursos — como recuperação textual e de imagens, fluxo condicional e memória de contexto — em uma única arquitetura voltada para o setor de energia. Além disso, destaca-se pelo foco em desempenho, com análise detalhada do tempo de resposta, e pela interface multimodal, que enriquece a experiência do usuário em ambientes operacionais.

III. Metodologia

O *chatbot* proposto foi desenvolvido com base em uma arquitetura que combina técnicas de *Embedding*, RAG e o uso de LLMs. Essa abordagem visa criar um sistema capaz de compreender perguntas formuladas em linguagem natural, recuperar as informações mais relevantes de uma base de dados textual e gerar respostas precisas e humanizadas.

Para estruturar o fluxo de operação do *chatbot*, foi utilizada a biblioteca *LangGraph*, que permite a construção de diagramas de estados para o controle do ciclo de vida da aplicação, desde a recepção da entrada do usuário até a geração da resposta final. A seguir, descrevem-se detalhadamente os principais componentes da aplicação desenvolvida.

A. Large Language Models (LLMs)

As LLMs são modelos de aprendizado profundo treinados com grandes volumes de dados textuais, com o objetivo de compreender e gerar linguagem natural de forma contextualizada e coerente. Esses modelos baseiam-se majoritariamente na arquitetura *transformer*, sendo capazes de realizar uma ampla gama de tarefas em Processamento de Linguagem Natural (PLN), como resposta a perguntas, geração de texto, sumarização e tradução automática [20].

Neste trabalho, optou-se pela utilização do modelo *LLaMA 3.3-70B*, desenvolvido pela *Meta AI*. Essa versão é composta por 70 bilhões de parâmetros e foi projetada para fornecer respostas mais precisas, fluentes e alinhadas às instruções do usuário. Este modelo se destaca pela eficiência no uso de memória e desempenho competitivo em tarefas de inferência, além de suportar instruções em linguagem natural com alta fidelidade semântica [21]. Sua robustez o torna uma escolha adequada para aplicações que exigem compreensão profunda do contexto e geração confiável de respostas.

A escolha deste modelo foi motivada por seu caráter *open-source*, que possibilita maior controle, personalização e integração ao sistema proposto. Além disso, o modelo demonstrou desempenho compatível com os requisitos de qualidade e confiabilidade esperados, apresentando uma boa relação entre custo computacional e precisão nas respostas. Tais características o tornaram uma escolha adequada neste contexto em relação a outros modelos disponíveis no momento.

Apesar de seu potencial, o uso de LLMs, como o *LLaMA 3.3-70B*, apresenta desafios importantes, como o risco de geração de informações imprecisas (hallucination) [20], a presença de vieses nos dados de treinamento e a alta demanda computacional [22]. Para minimizar esses riscos e aumentar a confiabilidade das respostas, adota-se a técnica de RAG, que será apresentada em detalhe nas seções seguintes.

B. Embedding

A técnica de *embedding* consiste em transformar textos em vetores numéricos de alta dimensionalidade, preservando relações semânticas no espaço vetorial [23]. Esse processo permite que palavras, sentenças ou documentos com significados semelhantes sejam representados por vetores próximos entre si, o que é fundamental para aplicações como busca semântica, sistemas de recomendação e classificação de textos [24].

A comparação entre esses vetores é geralmente realizada por meio da similaridade de cosseno, que avalia o ângulo entre dois vetores: quanto menor o ângulo, maior a similaridade entre os conteúdos representados [25]. Diferentemente de abordagens baseadas apenas em palavras-chave, os *embeddings* são capazes de capturar o significado contextual das frases, mesmo quando diferentes termos são utilizados, promovendo uma recuperação mais precisa e relevante das informações [26].

O fluxograma na Fig. 1 ilustra esse processo: documentos são lidos e divididos em trechos semânticos, que são então convertidos em vetores numéricos (*embeddings*). Esses vetores são armazenados em um banco vetorial, como o Milvus, permitindo buscas por similaridade semântica.

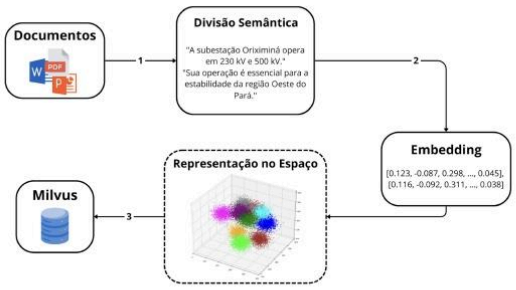


Fig. 1. Fluxograma do Processo de *Embedding*.

C. Recuperação Aumentada por Geração (RAG)

O sistema adota a abordagem RAG para melhorar a qualidade e a contextualização das respostas. Essa técnica combina a recuperação eficiente de trechos textuais relevantes com a capacidade de geração de linguagem natural de uma LLM [27].

No fluxo RAG, a pergunta do usuário é primeiro convertida em *embedding* e utilizada para buscar segmentos de documentos semanticamente similares na base vetorial. Os trechos recuperados são então incorporados como contexto para a LLM, que gera uma resposta coerente e informada, fundamentada nos dados recuperados [28].

Essa combinação permite que o sistema supere limitações comuns a modelos geradores puros, que podem inventar informações ou responder de forma genérica, ao incorporar diretamente conhecimento extraído da base de dados indexada. O RAG, portanto, melhora a precisão, relevância e atualidade das respostas geradas [29]. A Fig. 2 ilustra esse fluxo de forma esquemática, apresentando as etapas que compõem a arquitetura RAG utilizada no sistema.

D. Geração da Base de Dados

A base de dados do sistema foi construída a partir da ingestão de documentos localizados em um diretório, utilizando o componente *DirectoryLoader* da biblioteca *LangChain*. Esse componente permite o carregamento de arquivos com diferentes formatos, padronizando o conteúdo para processamento posterior. A coleção utilizada era composta por 10 documentos técnicos contendo orientações detalhadas sobre os procedimentos que devem ser executados pelos controladores em diferentes cenários operacionais,

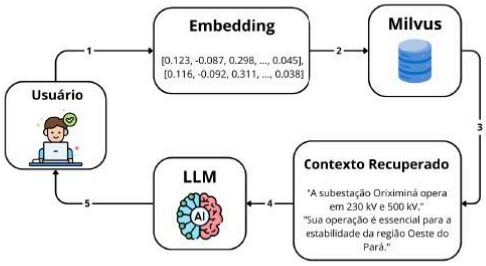


Fig. 2. Fluxograma de funcionamento do RAG.

incluindo etapas de manobras, protocolos de segurança e instruções para situações de emergência.

Para viabilizar uma recuperação eficiente e contextualizada, os documentos foram segmentados em unidades semânticas menores, como parágrafos ou sentenças. Essa segmentação tem como objetivo estruturar a base em trechos com coesão temática suficiente para garantir relevância no momento da busca.

Cada segmento textual foi convertido em um vetor numérico de alta dimensionalidade, ou *embedding*, utilizando o modelo '*paraphrase-multilingual-mpnet-base-v2*' da biblioteca *SentenceTransformer*.

E. Base de Dados Vetorial

Os *embeddings* gerados foram armazenados em uma base vetorial construída com o *Milvus*, uma biblioteca especializada no gerenciamento e recuperação eficiente de dados vetoriais em larga escala. Esta possibilita buscas rápidas e escaláveis por similaridade semântica, sendo fundamental para a etapa de recuperação de informações do sistema [30].

F. Processamento e Indexação de Imagens

Além do texto, o sistema integra informações visuais para enriquecer as respostas. Cada imagem da base foi associada a um rótulo textual descritivo, organizado em formato JSON contendo o texto do rótulo e o diretório do arquivo correspondente. Esses rótulos também foram convertidos em *embeddings* utilizando o mesmo modelo semântico '*paraphrase-multilingual-mpnet-base-v2*', garantindo compatibilidade no espaço vetorial com os embeddings textuais.

Os vetores dos rótulos foram indexados em uma segunda tabela na base *Milvus*, dedicada exclusivamente à recuperação por similaridade semântica das imagens. Durante a inferência, essa base é consultada para sugerir imagens relevantes que complementam o conteúdo da resposta textual, promovendo uma experiência multimodal ao usuário.

G. Arquitetura Lógica e Fluxo de Execução com LangGraph

Para coordenar o fluxo de execução do *chatbot*, foi utilizada a biblioteca *LangGraph*, que permite modelar processos computacionais por meio de grafos direcionados. Cada nó representa uma operação funcional — como a recepção da pergunta, busca de contexto ou geração da resposta — enquanto as arestas definem transições condicionais entre essas etapas [31].

Esse modelo gráfico oferece flexibilidade para definir fluxos condicionais, ciclos e ramificações, essencial em sistemas de linguagem natural onde diferentes tipos de perguntas e dados disponíveis demandam tratamentos diferenciados. O fluxo é organizado nos seguintes nós principais:

- **Recuperar Contexto:** recebe a pergunta do usuário, converte-a em *embedding* com o modelo '*paraphrase-multilingual-mpnet-base-v2*' e realiza busca por similaridade na base textual. Os segmentos mais relevantes são armazenados como contexto para geração da resposta.
- **Gerar Resposta:** utiliza a pergunta original e o contexto recuperado para alimentar uma LLM que, guiada por *prompts* pré-definidos, gera uma resposta textual coerente, contextualizada e natural.
- **Recuperar Imagens:** recebe a resposta gerada pela LLM, converte-a em *embedding* semântico e consulta a base vetorial de rótulos de imagens para identificar e retornar a imagem mais relevante para enriquecer a resposta.

Além disso, blocos condicionais no grafo avaliam a necessidade de incluir imagens na resposta (verificando presença de palavras-chave na pergunta) e a relevância do contexto recuperado para decidir entre gerar uma resposta elaborada ou apresentar uma mensagem padrão caso não haja dados suficientes.

Outro diferencial da biblioteca utilizada é seu mecanismo interno de memória de estado, capaz de registrar os estados de todos os nós do grafo a cada iteração. Esse recurso permite a construção de um histórico de conversas, possibilitando que o chatbot tenha acesso às interações anteriores para formular respostas mais contextualizadas em novos diálogos [31]. Como resultado, a experiência do usuário é aprimorada, uma vez que o sistema pode manter a coerência e a continuidade durante sessões de conversa prolongadas.

A Fig. 3 ilustra o fluxo construído para o chat.

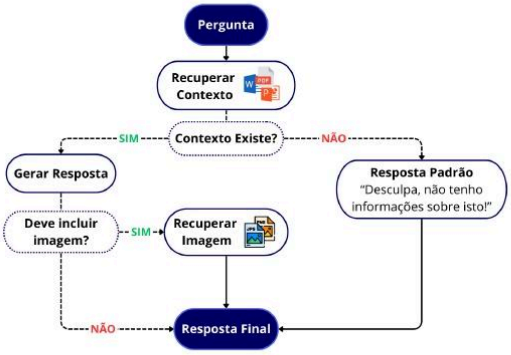


Fig. 3. Fluxograma de funcionamento do Chat.

H. Interface Gráfica

Com o objetivo de aprimorar a experiência do usuário e assegurar uma apresentação interativa, responsiva e de fácil utilização, o *front-end* do sistema de chat foi desenvolvido utilizando HTML, CSS e JavaScript, enquanto o *back-end* foi implementado com o *framework Flask*. Essa abordagem

permitiu a criação de uma aplicação web com maior controle sobre a interface, garantindo flexibilidade na personalização e integração dos componentes visuais com a lógica do sistema. Além disso, o chat conta com a possibilidade de criar múltiplas conversas, e o histórico é garantido pelo armazenamento no cache do navegador, permitindo que o usuário retome interações anteriores de forma rápida e contínua.

I. Estrutura Geral do ChatBot

Na Fig. 4 é possível visualizar a função de cada técnica e biblioteca empregada na criação do *chatbot*. A arquitetura é composta por quatro etapas principais. Inicialmente, o usuário interage com a aplicação por meio da interface desenvolvida, onde insere sua pergunta. Essa solicitação é processada pelo módulo *LangGraph*, responsável por coordenar o fluxo de informações e realizar a recuperação de contexto. Para isso, o LangGraph consulta a base vetorial *Milvus*, que armazena representações semânticas de documentos previamente processados. A vetorização desses documentos (PDF, *Word*, *PowerPoint*, entre outros) é realizada por meio do modelo *Sentence Transformer*, possibilitando buscas semânticas eficientes. Com o contexto recuperado, o *LangGraph* encaminha as informações relevantes a um LLM, que é responsável pela geração da resposta textual. Por fim, essa resposta é enviada de volta à interface, onde é exibida ao usuário de forma clara e objetiva.

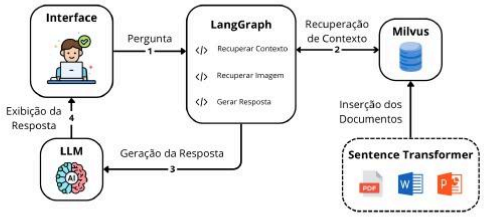


Fig. 4. Estrutura do Chat.

IV. Resultados e Discussões

A implementação do *chatbot* resultou em uma interface interativa e funcional, voltada para facilitar o acesso às informações operacionais de uma empresa do setor de energia. A Fig. 5 ilustra a interface final do sistema, permitindo que o usuário realize consultas em linguagem natural, recebendo respostas precisas e contextualizadas com base nos dados indexados.

Fig. 5. Interface final do Chat.

Para avaliar a precisão e a capacidade de compreensão do sistema, foi realizada uma consulta sobre as conexões da Subestação Oriximiná. O chatbot respondeu listando corretamente todas as linhas de transmissão (LTs) associadas à subestação, com suas respectivas tensões e destinos, além de fornecer uma explicação complementar sobre as interligações com as subestações Juruti, Jurupari e Silves.

Ao comparar a resposta gerada com o conteúdo original, observa-se uma correspondência exata com os dados estruturados: todas as seis LTs foram identificadas corretamente, os níveis de tensão foram mencionados com precisão. Além disso, o sistema devolveu ao usuário um diagrama da rede base que, embora não estivesse presente nos dados originais, foi incluído como elemento visual de apoio. Este recurso multimodal oferece ao usuário uma representação espacial da rede de transmissão na região do Pará.

Esse resultado evidencia a capacidade do *chatbot* em interpretar a intenção do usuário, recuperar e organizar informações técnicas com precisão, e apresentar a resposta de forma clara e visualmente compreensível. A integração entre a linguagem natural e o componente gráfico reforça a eficácia da abordagem adotada, sobretudo em contextos operacionais que exigem interpretação rápida e precisa de dados complexos.

Os testes também demonstraram que a integração da técnica RAG com o modelo escolhido atuou como um mecanismo eficaz para mitigar a geração de respostas imprecisas. Ao fornecer à LLM trechos relevantes previamente extraídos da base vetorial, o sistema reduziu significativamente a ocorrência de *hallucinations* e aumentou a confiabilidade das respostas. Essa integração direta entre recuperação e geração mostrou-se a principal estratégia para assegurar precisão técnica nas respostas.

Além disso, foram realizados testes para medir o tempo médio de resposta do chatbot. Para isso, foram feitas trinta perguntas representativas da base de informações do sistema, a qual, durante os testes, era composta por 10 documentos técnicos contendo orientações detalhadas sobre os procedimentos que devem ser executados pelos controladores em diferentes cenários operacionais, incluindo etapas de manobras, protocolos de segurança e instruções para situações de emergência. O objetivo foi avaliar a performance em diferentes etapas do processamento das solicitações. A Tabela I apresenta os tempos médios obtidos para cada etapa principal do fluxo de resposta.

TABLE I
Tempos médios por nó após 30 iterações

Etapa	Tempo Médio (s)
Recuperar Contexto	0,1528
Condicional Contexto	0,0000
Gerar Resposta	1,1685
Condicional Imagem	0,0001
Recuperar Imagens	0,3043
Tempo médio total	**1,6257**

Observa-se que a maior parte do tempo de processamento está concentrada na etapa de geração da resposta, que corresponde em média a cerca de 1,17 segundos por interação. A recuperação de imagens também impacta o tempo total, com cerca de 0,30 segundos. As etapas condicionais, por sua vez,

têm tempos desprezíveis, indicando que a lógica condicional não adiciona atrasos significativos.

No geral, o tempo médio total estimado para o sistema responder a uma pergunta é de aproximadamente 1,63 segundos, um valor considerado satisfatório para garantir uma experiência fluida e interativa para o usuário. Esse desempenho indica que o chatbot é capaz de processar as informações e gerar respostas em tempo real, minimizando a sensação de espera.

V. Conclusões

Tendo isso em vista, o chatbot desenvolvido se consolida não apenas como um assistente de consulta, mas como uma ferramenta estratégica para a gestão da informação em tempo real, contribuindo para uma tomada de decisão mais ágil, precisa e embasada, especialmente em cenários críticos de operação no setor elétrico.

A fluidez do fluxo de interação, viabilizada pela biblioteca LangGraph, proporciona uma experiência intuitiva, com respostas rápidas, claras e adaptadas à linguagem técnica dos usuários. Essa característica representa um avanço significativo em relação aos métodos tradicionais de consulta a documentos, muitas vezes dispersos em sistemas internos ou arquivos manuais.

Do ponto de vista de desempenho, o sistema demonstrou robustez na recuperação semântica, mesmo diante de perguntas formuladas com termos distintos dos presentes na base de dados. Esse resultado é atribuído à integração eficaz das técnicas de *embedding* e RAG, que garantem respostas mais relevantes, contextualizadas e informativas.

Assim, conclui-se que a arquitetura proposta representa um avanço tecnológico aplicável a contextos industriais, com potencial de expansão para outras áreas mediante a incorporação de novas fontes de dados e funcionalidades, como aquelas previstas nos trabalhos futuros.

VI. Trabalhos Futuros

Como trabalho futuro, propõe-se a implementação do $chatbot$ em ambiente local, por meio da plataforma *Ollama*, com o objetivo de garantir maior controle sobre a execução do modelo, aumentar a segurança dos dados e permitir o uso de informações sensíveis sem risco de vazamentos.

Além disso, busca-se ampliar o acesso a dados operacionais por meio da integração do chatbot com o banco de dados corporativo. Para isso, será necessário desenvolver uma camada de integração entre a LLM e o banco de dados relacional, permitindo que comandos em linguagem natural sejam automaticamente convertidos em consultas SQL, viabilizando a recuperação dinâmica e precisa de informações diretamente do sistema de dados da empresa.

VII. Agradecimentos

Os autores agradecem ao Programa de Educação Tutorial de Engenharia Elétrica (PET-EE) da Universidade Federal do Pará (UFPA) e a concessionária Equatorial Energia Pará.

References

[1] N. R. Delbianco and M. L. P. Valentim, "Sociedade da informação e as mídias sociais no contexto da comunicação científica," *AtoZ: novas práticas em informação e conhecimento*, vol. 11, pp. 1–11, 2022.

[2] E. F. T. Hernandez and N. K. A. de Toledo, "Crimes cibernéticos: seus efeitos revolucionários diante de uma legislação em constante evolução," *Revista Jurídica da UniFil*, vol. 17, no. 17, pp. 72–84, 2021.

[3] L. P. M. Gäal and M. S. Martins, "Acesso aberto no contexto da pesquisa em ciência da informação," *Transinformação*, vol. 34, p. e220016, 2022.

[4] F. V. Perfetto, S. G. d. O. Reis, and F. C. Paletta, "Gestão da informação digital caminhos possíveis," *RDBCI: Revista Digital de Biblioteconomia e Ciência da Informação*, vol. 21, p. e023005, 2023.

[5] A. L. Albertin and R. M. de Moura Albertin, "Transformação digital: gerando valor para o" novo futuro"," *Gv-Executivo*, vol. 20, no. 1, pp. 26–29, 2021.

[6] M. de Oliveira Morais, P. L. d. O. C. Neto, O. S. dos Santos, A. P. Cardoso Jr, and J. B. Sacomano, "A evolução da qualidade na indústria 4.0," *Research, Society and Development*, vol. 9, no. 10, pp. e3 929 108 634–e3 929 108 634, 2020.

[7] V. F. de Oliveira, "Arquitetura para integração e análise de dados baseada no padrão indústria 4.0." 2023.

[8] M. J. P. de Antônio, J. P. Alves, D. T. do Nascimento, and M. C. de Moura Garrido, "Automação logística, ia e indústria 4.0: Considerações éticas em um cenário de inovação tecnológica," *Mobicities-Journal of Urban Mobility, Logistics and Sustainable Smart Cities*, vol. 2, no. 1, 2025.

[9] L. G. da Costa Neto and F. C. de Campos, "Oportunidades de aplicações de business intelligence no contexto da indústria 4.0: Revisão sistemática da literatura 2015-2020," *Exacta*, vol. 21, no. 2, pp. 503–519, 2023.

[10] A. A. d. A. Freire, "Ifpbbot: um chatbot acadêmico," B.S. thesis, 2024.

[11] G. Pallin, D. Lemes, H. M. Ali, M. Coelho, M. Rosa, R. Christino, A. C. Senger, and L. A. Mathias, "Vacabot: Um chatbot que emprega linguagem natural no ambiente de viagens," in *Anais do Congresso Brasileiro de Iniciação Cientıfica*, vol. 1, no. 2, 2024, pp. 77–82.

[12] A. S. MENEGUELLO, B. F. MARQUES, E. G. G. RODRIGUES, and W. d. B. FERREIRA, "Chatbot educacional: desenvolvimento de um chatbot para estudantes e educadores," 2023.

[13] V. A. Lugli and J. de Lucca Filho, "O uso do chatbot para a excelência em atendimento," *Revista Interface Tecnológica*, vol. 17, no. 1, pp. 205–218, 2020.

[14] A. Martins and P. A. Paiva, "Inteligência artificial nas empresas: estudo de caso do chatbot lz," *Brazilian Journal of Management and Innovation (Revista Brasileira de Gestão e Inovação)*, vol. 12, no. 1, 2025.

[15] C. G. Larussa Filho, J. H. G. Borges, C. H. M. Porta, and F. Florian, "Desenvolvimento de chatbot para consulta de informações acadêmicas do aluno," *RECIMA21-Revista Cientıfica Multidisciplinar-ISSN 2675-6218*, vol. 3, no. 12, pp. e3 122 364–e3 122 364, 2022.

[16] R. Parkar, Y. Payare, K. Mithari, J. Nambiar, and J. Gupta, "Ai and web-based interactive college enquiry chatbot," pp. 1–5, 2021.

[17] H.-C. Lee, K. Hung, G. M.-T. Man, R. Ho, and M. Leung, "Development of an rag-based llm chatbot for enhancing technical support service," pp. 1080–1083, 2024.

[18] L. Lu, Y. Zhou, Z. Wang, J. Liu, H. Jin, and Q. Guo, "Power system simulation intelligent assistant: A chatbot integrating large language models and domain models," pp. 249–253, 2024.

[19] N. S. d. Oliveira, "Desenvolvimento de um assistente chatbot inteligente para instalações elétricas baseado em modelo de linguagem grande (llm)," 2024.

[20] S. Minaee, T. Mikolov, N. Nikzad, M. Chenaghlu, R. Socher, X. Amatriain, and J. Gao, "Large language models: A survey," *arXiv preprint arXiv:2402.06196*, 2024. [Online]. Available: https://arxiv.org/abs/2402.06196

[21] Meta AI, "Introducing llama 3: Open foundation and instruction-tuned models," https://ai.meta.com/llama/, 2024, acesso em: 21 maio 2025.

[22] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, "On the dangers of stochastic parrots: Can language models be too big?" in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*. Virtual Event, Canada: ACM, 2021, pp. 610–623.

[23] T. Rodrigues *et al.*, "Portuguese language models and word embeddings: Evaluating on semantic similarity tasks," in *Experimental IR Meets Multilinguality, Multimodality, and Interaction*. Springer, 2020, pp. 295–307.

[24] L. Tang, "UZH at SemEval-2020 task 3: Combining BERT with WordNet sense embeddings to predict graded word similarity changes," in *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, A. Herbelot, X. Zhu, A. Palmer, N. Schneider, J. May, and E. Shutova, Eds. Barcelona (online): International Committee

for Computational Linguistics, Dec. 2020, pp. 166–170. [Online]. Available: https://aclanthology.org/2020.semeval-1.19/

[25] J. Zhou *et al.*, "Bio-inspired structure identification in language embeddings," *arXiv preprint arXiv:2009.02459*, 2020. [Online]. Available: https://arxiv.org/abs/2009.02459

[26] D. Novotný *et al.*, "Text classification with word embedding regularization and soft similarity measure," *arXiv preprint arXiv:2003.05019*, 2020. [Online]. Available: https://arxiv.org/abs/2003.05019

[27] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel *et al.*, "Retrieval-augmented generation for knowledge-intensive nlp tasks," *arXiv preprint arXiv:2005.11401*, 2020.

[28] Y. Huang and J. Huang, "A survey on retrieval-augmented text generation for large language models," *arXiv preprint arXiv:2404.10981*, 2024.

[29] Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, M. Wang, and H. Wang, "Retrieval-augmented generation for large language models: A survey," *arXiv preprint arXiv:2312.10997*, 2023.

[30] J. Wang, X. Yi, R. Guo, H. Jin, P. Xu, S. Li, X. Wang, X. Guo, C. Li, X. Xu, K. Yu, Y. Yuan, Y. Zou, J. Long, Y. Cai, Z. Li, Z. Zhang, Y. Mo, J. Gu, R. Jiang, Y. Wei, and C. Xie, "Milvus: A purpose-built vector data management system," in *Proceedings of the 2021 International Conference on Management of Data (SIGMOD '21)*. Virtual Event, China: ACM, 2021, pp. 2614–2627.

[31] LangChain, Inc., "Langgraph," 2025, a framework open-source para desenvolvimento de agentes de IA adaptáveis e confiáveis.