josif 2024
16ª JORNADA CIENTÍFICA E TECNOLÓGICA
13º SIMPÓSIO DE PÓS-GRADUAÇÃO
DO IFSULDEMINAS
ISSN: 2319-0124
REALIZAÇÃO INSTITUTO FEDERAL Sul de Minas Gerais
APOIO CNPq FAPEMIG

ZENOAI: Otimização de Suporte Técnico e Treinamento Corporativo com RAG e LLM

Diogo H. da SILVA[1]; Paulo C. Dos SANTOS[2]

RESUMO

Com o aumento da rotatividade e a complexidade dos processos corporativos, aprimorar o suporte técnico e o treinamento interno torna-se essencial. O ZenoAI é uma solução que utiliza Retrieval-Augmented Generation (RAG) e Large Language Models (LLM) para gerar respostas precisas e contextuais com base nos processos específicos da empresa. Integrando informações internas e automatizando o treinamento, o ZenoAI melhora a eficiência e a eficácia do suporte técnico, adaptando-se continuamente às necessidades da organização. Ao adotar essa abordagem, não dependemos apenas de fontes de dados externas, mas garantimos a gestão local dos dados, beneficiando-se da estrutura do RAG. Este artigo detalha a arquitetura do sistema, a aplicação das tecnologias e a metodologia desenvolvida, demonstrando como o ZenoAI contribui para um suporte técnico mais ágil e personalizado.

Palavras-chave: Retrieval-Augmented Generation; Large Language Models; Suporte Técnico; Treinamento Corporativo.

1. INTRODUÇÃO

Com o aumento da rotatividade e a crescente complexidade dos processos corporativos, melhorar a eficiência do suporte técnico e o treinamento interno tornou-se uma prioridade para muitas organizações. Tecnologias emergentes, como a Retrieval-Augmented Generation (RAG) e os Large Language Models (LLMs), oferecem novas abordagens para enfrentar esses desafios e otimizar o gerenciamento do conhecimento.

A Retrieval-Augmented Generation (RAG) representa uma evolução significativa em modelos de linguagem, combinando a geração de texto com a recuperação de informações externas. Diferentemente dos modelos de linguagem tradicionais, que se baseiam exclusivamente nos dados de treinamento, o RAG permite a integração de dados em tempo real, melhorando a precisão e a relevância das respostas geradas (GAO *et al.*, 2023).

O sistema RAG é composto por três componentes principais: o recuperador, responsável por buscar e recuperar informações relevantes; o gerador, que utiliza essas informações para criar respostas; e o mecanismo de ampliação, que atualiza continuamente o conhecimento disponível (GAO *et al.*, 2023).

A integração de RAG e LLMs no suporte técnico e treinamento corporativo pode transformar a forma como as empresas gerenciam o conhecimento e formam seus funcionários. A aplicação do RAG permite que o sistema se adapte às necessidades específicas da organização, oferecendo respostas contextualizadas e precisas, enquanto garante maior segurança ao reduzir a dependência de

[1]Discente, IFSULDEMINAS – *Campus* Muzambinho. E-mail: diogo1.silva@alunos.ifsuldeminas.edu.br
[2]Orientador, IFSULDEMINAS – *Campus* Muzambinho. E-mail: paulo.santos@muz.ifsuldeminas.edu.br.

fontes externas (AHMED, 2024).

Este artigo explora a implementação do ZenoAI, detalhando a arquitetura do sistema, a aplicação das tecnologias RAG e LLM, e os benefícios associados à utilização dessas soluções avançadas no ambiente corporativo.

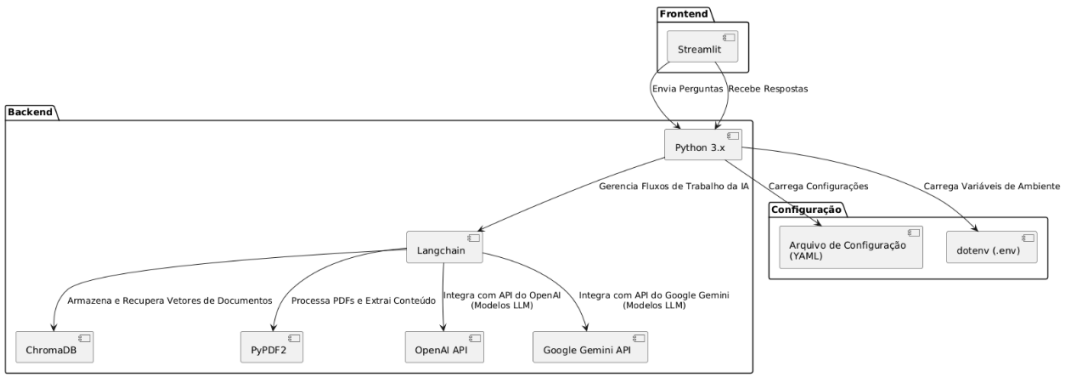
2. MATERIAL E MÉTODOS

O desenvolvimento do ZenoAI seguiu uma abordagem estruturada com o uso do Jira para gerenciamento de tarefas e do GitHub como repositório central. O ambiente foi configurado em um notebook com Intel Core i7, 20 GB de RAM, e sistema Windows 10, utilizando o Visual Studio Code para edição de código.

A implementação utilizou as bibliotecas Streamlit, para a interface interativa, e LangChain, para integração dos modelos de linguagem com bases de dados. A extração de texto de documentos foi feita com PyPDF2, e o armazenamento vetorial foi gerido pelo Chroma. Modelos de linguagem, fornecidos por APIs da OpenAI e Google Gemini, suportaram a geração e recuperação de informações baseadas nos documentos processados.

A arquitetura do sistema foi projetada para otimizar a interação entre esses componentes, e uma representação gráfica desta arquitetura, gerada com o PlantUML WebServer.

Figura 1: Arquitetura do sistema ZenoAI.



Fonte: dos autores.

A Figura 1 representa o diagrama da arquitetura do ZenoAI, ilustrando as conexões e interações entre os principais componentes: a interface do usuário, a integração com APIs de modelos de linguagem (LLMs) e o gerenciamento do banco de dados vetorial..

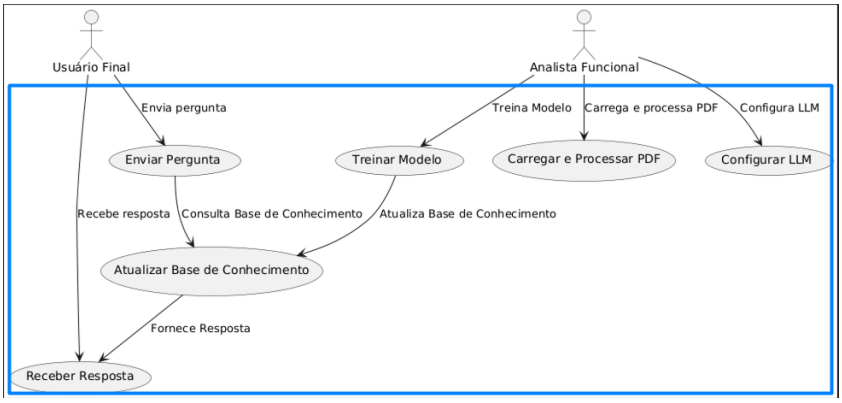
3. RESULTADOS E DISCUSSÃO

A aplicação ZenoAI foi projetada para responder a perguntas funcionais e técnicas sobre o novo sistema da empresa, utilizando o contexto fornecido pelo documento PDF como base de conhecimento. Para avaliar sua eficácia, foram realizados testes com usuários reais, que foram selecionados e submetidos a perguntas sobre o sistema.

Os testes demonstraram que o ZenoAI é capaz de fornecer respostas precisas e contextualmente relevantes. A taxa de acerto foi superior a 85%, com um baixo nível de alucinação, confirmando a efetividade do sistema na compreensão e na resposta às perguntas baseadas no contexto do documento.

A interface desenvolvida com Streamlit facilitou a interação dos usuários com o sistema, tornando o processo intuitivo e eficiente.

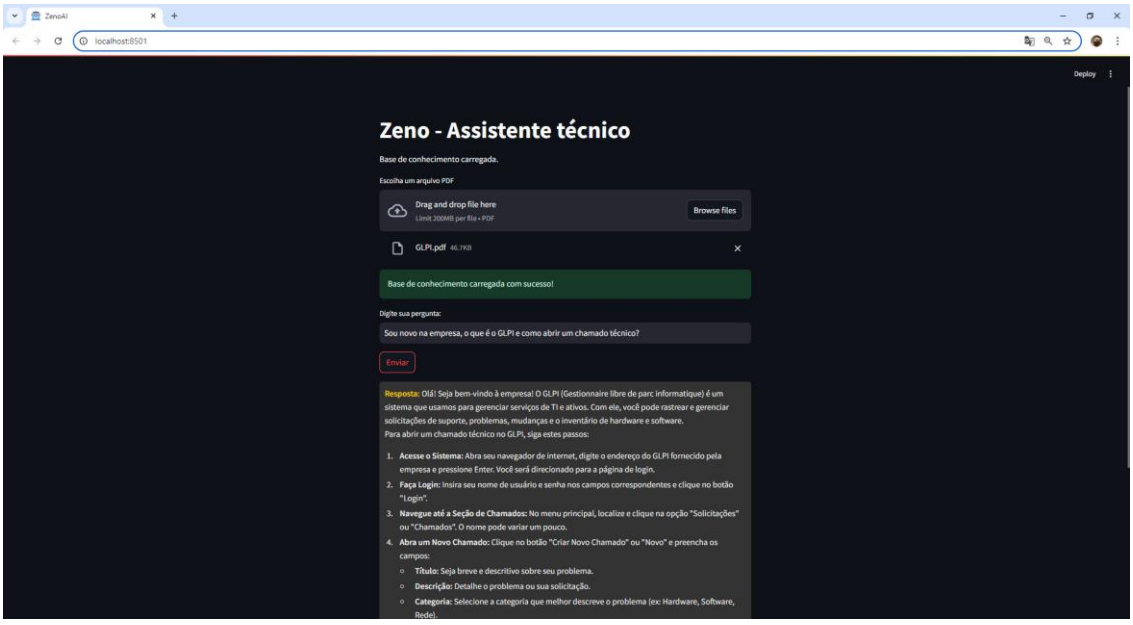
Figura 2: Diagrama de caso de uso do sistema ZenoAI.



Fonte: dos autores.

A Figura 2 ilustra como o usuário interage com as funcionalidades do sistema, apresentando o fluxo de suporte técnico automatizado.

Figura 3: Interface do sistema ZenoAI em uso.



Fonte: dos autores.

A Figura 3 apresenta uma imagem do sistema em uso, demonstrando como os usuários podem fazer perguntas e obter respostas contextuais, baseadas nos documentos corporativos, através da interface interativa da aplicação.

Esses resultados confirmam que o ZenoAI oferece um suporte técnico eficaz, ajudando na

adaptação dos novos funcionários ao sistema e fornecendo uma ferramenta confiável para a resolução de dúvidas técnicas e funcionais.

4. CONCLUSÃO

O projeto ZenoAI demonstrou sucesso na criação de um assistente técnico para responder perguntas sobre o novo sistema da empresa. Utilizando um documento PDF como contexto, a aplicação forneceu respostas precisas tanto para perguntas funcionais quanto técnicas.

Os testes com usuários reais confirmaram a eficácia do sistema, destacando seu baixo nível de alucinação e alta relevância das respostas. A integração das tecnologias OpenAI, Google Gemini, LangChain, e Chroma, junto com a interface desenvolvida em Streamlit, resultou em uma solução eficiente e intuitiva para suporte técnico. Assim, o ZenoAI atingiu seus objetivos, oferecendo uma ferramenta eficaz para a compreensão do novo sistema pela equipe.

REFERÊNCIAS

AHMED, S. *What is Retrieval-Augmented Generation (RAG) in LLM and How it works?*. Medium, 22 abr. 2024. Disponível em: https://medium.com/@sahin.samia/what-is-retrieval-augmented-generation-rag-in-llm-and-how-it-works-a8c79e35a172. Acesso em: 25 ago. 2024.

GAO, Y. *Retrieval-Augmented Generation for Large Language Models: A Survey*. Disponível em: https://arxiv.org/abs/2312.10997. Acesso em: 25 ago. 2024.

LANGCHAIN. LangChain: The Framework for Developing Applications Powered by Language Models. Disponível em: https://www.langchain.com. Acesso em: 17 ago. 2024.

STREAMLIT. Streamlit: The Fastest Way to Build and Share Data Apps. Disponível em: https://www.streamlit.io. Acesso em: 17 ago. 2024.