ASSISTENTE VIRTUAL PARA SUPORTE ACADÊMICO: IMPLEMENTAÇÃO NO CURSO DE LICENCIATURA EM COMPUTAÇÃO COM IA GENERATIVA E RAG

Ronald Ruan Pereira Soares[1], Fábio Emanuell Abreu Cardoso[2], Julio Cesar Santos Sousa[3], Nicolas Heitor Feitosa Costa[4], Ramasio Ferreira de Melo[5]

[1]Graduado pelo Curso de Licenciatura em Computação do Instituto Federal de Educação, Ciência e Tecnologia do Tocantins- IFTO, ronald.soares@estudante.ifto.edu.br
[2]Graduando do Curso de Licenciatura em Computação do Instituto Federal de Educação, Ciência e Tecnologia do Tocantins- IFTO, fabio.cardoso2@estudante.ifto.edu.br
[3]Graduando do Curso de Licenciatura em Computação do Instituto Federal de Educação, Ciência e Tecnologia do Tocantins- IFTO, julio.sousa6@estudante.ifto.edu.br
[4]Graduando do Curso de Licenciatura em Computação do Instituto Federal de Educação, Ciência e Tecnologia do Tocantins- IFTO, nicolas.costa@estudante.ifto.edu.br
[5]Docente do Curso Superior de Licenciatura em Computação – IFTO. Orientador. e-mail: ramasiomelo@ifto.edu.br

1 INTRODUÇÃO

O uso de Inteligência Artificial (IA) na educação tem se expandido rapidamente, impulsionado por grandes modelos de linguagem natural. Agentes de IA e chatbots têm sido utilizados em atividades pedagógicas e na gestão institucional para otimizar tarefas repetitivas e melhorar a comunicação com os estudantes (Lee et al., 2019; Li, 2023; Dias, 2022). Estudos demonstram que esses chatbots oferecem interatividade, feedback em tempo real e assistência personalizada, contribuindo para a eficiência de processos pedagógicos e administrativos (Pérez; Daradoumis; Puig, 2020; Dempore et al., 2023). Pesquisas recentes também destacam o impacto positivo na organização interna das instituições, na redução da carga de trabalho dos servidores e na ampliação do acesso à informação (Ilieva et al., 2023; Mendoza et al., 2022).

Apesar dos avanços, muitas instituições ainda enfrentam desafios na implementação de assistentes virtuais adaptados às suas realidades. Soluções genéricas que se limitam a respostas a perguntas frequentes (FAQ) não consideram o contexto institucional e as demandas específicas de estudantes e servidores. Além disso, há uma escassez de estudos que exploram a aplicação da técnica de Recuperação com Geração Aumentada (RAG) em ambientes educacionais, especialmente em instituições públicas de ensino superior.

Para preencher essa lacuna, este artigo apresenta o desenvolvimento de um assistente virtual para o curso de Licenciatura em Computação do IFTO – Campus Araguatins. A solução foi criada com base no modelo GPT-4-mini, integrada ao WhatsApp e apoiada na técnica RAG, que combina a busca por informações relevantes com a geração de respostas contextualizadas. O sistema utiliza a plataforma Supabase para armazenar dados vetorizados e acessa documentos institucionais do Google Drive, garantindo respostas atualizadas sobre temas como TCC, estágio, regulamentos e o Projeto Pedagógico de Curso (PPC). As principais contribuições são: (i) a implementação de um assistente com IA generativa, (ii) a integração da arquitetura RAG para consulta a informações acadêmicas e (iii) a escolha do WhatsApp como canal de interação, promovendo acessibilidade e inclusão digital.

2 OBJETIVO

Este trabalho tem como objetivo principal desenvolver e implementar um assistente virtual para o curso de Licenciatura em Computação do Instituto Federal do Tocantins - Campus Araguatins, utilizando a técnica de Recuperação com Geração Aumentada (RAG) em conjunto com modelos de linguagem em grande escala (LLMs), a fim de automatizar o atendimento e fornecer respostas precisas e contextualizadas a dúvidas frequentes da comunidade acadêmica.

3 MATERIAL E MÉTODOS

A presente pesquisa adota uma abordagem metodológica descritiva e aplicada, focada no desenvolvimento de um assistente virtual como suporte acadêmico para o curso de Licenciatura em Computação do IFTO - Campus Araguatins. A metodologia seguiu um fluxo de engenharia de software, tendo como base a arquitetura de Recuperação com Geração Aumentada (RAG), que combina a busca por documentos relevantes com a geração de respostas contextualizadas por meio de modelos de linguagem de grande escala (LLMs).

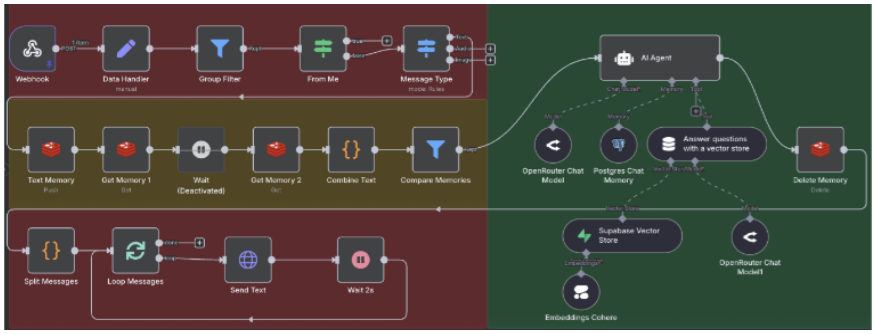
Para adaptar o LLM ao contexto do curso, uma base de conhecimento foi construída a partir de 36 documentos institucionais, como Projetos Pedagógicos de Curso (PPCs), atas e regulamentos. Os arquivos, armazenados no Google Drive, passaram por uma limpeza manual para remoção de cabeçalhos e formatação de tabelas, sendo em seguida convertidos para o formato Markdown (.md). O processamento utilizou a ferramenta de automação n8n, segmentando o texto em blocos de 1.700 caracteres com sobreposição de 450, para preservar a coesão. Cada bloco foi transformado em um vetor semântico pelo modelo de embeddings Cohere v3.0 Multilingual e armazenado em uma base de dados vetorial no Supabase.

O assistente virtual foi implementado para interagir com os usuários via WhatsApp. Ao receber uma pergunta, o sistema a converte em um vetor, busca os trechos mais relevantes na base de dados e os utiliza como contexto para o modelo GPT-4-mini, que gera a resposta. A avaliação do sistema foi realizada por meio de testes manuais, comparando as respostas a documentos oficiais para verificar sua coerência e precisão. A pesquisa não envolveu a coleta de dados de seres humanos.

4 RESULTADOS E DISCUSSÃO

O assistente virtual foi implementado com sucesso e integrado ao WhatsApp, permitindo que os usuários realizem perguntas em linguagem natural e recebam respostas em tempo real. A Figura 1 ilustra o fluxo completo do assistente, desde a recepção da mensagem por meio de um Webhook até o processamento e envio da resposta. O sistema extrai as informações do usuário, valida a entrada e, após essa filtragem, armazena temporariamente o contexto da conversa. Em seguida, a pergunta é processada por um modelo de linguagem baseado na arquitetura de Recuperação com Geração Aumentada (RAG), que permite gerar respostas coerentes e contextualizadas, garantindo um atendimento dinâmico, eficiente e escalável.

Figura 1 - Fluxo completo do assistente virtual.

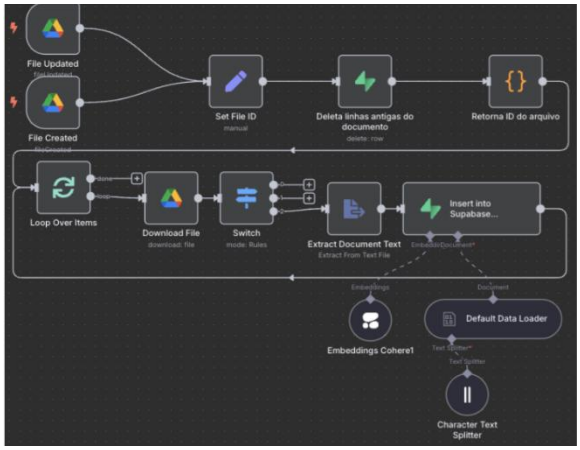


Elaborado pelos autores (2025).

Além disso, foi desenvolvido um fluxo automatizado para a vetorização de documentos institucionais do Google Drive, conforme demonstrado na Figura 2. As configurações adotadas para a estrutura de vetorização foram definidas com base em testes que buscaram equilibrar desempenho

e compreensão semântica, com o tamanho do chunk fixado em 1700 tokens e um overlap de 450 para preservar o contexto. A escolha do modelo de embeddings Cohere v3.0 Multilingual (dimensão 1024) se deu por sua alta qualidade na representação semântica em português, o que favorece futuras expansões.

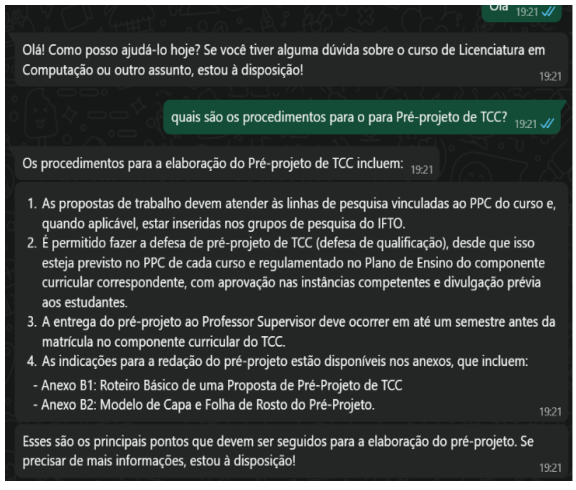
Figura 2 - Fluxo de automação no n8n responsável pela vetorização de documentos.



Elaborado pelos autores (2025).

Durante os testes preliminares, o sistema demonstrou boa capacidade de interpretar questões específicas relacionadas ao curso de Licenciatura em Computação. O assistente respondeu corretamente a perguntas frequentes, como os procedimentos para elaboração do pré-projeto de TCC, listando as diretrizes do curso e os anexos corretos (Anexo B1 e B2), como pode ser visto na Figura 3. Da mesma forma, respondeu com precisão sobre as disciplinas do oitavo período e a carga horária total do curso, que é de 280 horas. Esses resultados confirmam a capacidade do agente em compreender a intenção do usuário e recuperar informações precisas e contextualizadas do repositório de documentos, evidenciando a eficácia da abordagem baseada em RAG.

Figura 3 - Resposta do assistente sobre os procedimentos para elaboração do pré-projeto de TCC.



Elaborado pelos autores (2025).

5 CONSIDERAÇÕES FINAIS

A pesquisa demonstrou uma solução inovadora para o atendimento acadêmico. A proposta se destaca por fornecer respostas precisas e contextualizadas, usando dados reais do curso, o que a diferencia de assistentes genéricos. Apesar disso, o sistema apresenta limitações, como dificuldades com perguntas ambíguas e a dependência da qualidade dos dados e da restrição de tokens dos modelos de linguagem.

Sendo assim, esses desafios apontam para a necessidade de futuros trabalhos que incluam testes em outros contextos e formatos de documentos para aumentar a robustez da solução. É fundamental, também, adotar indicadores de impacto para mensurar o uso do sistema, o tempo médio de resposta e a satisfação dos usuários. Por fim, recomenda-se a expansão do projeto com a implementação de um servidor local para rodar um LLM e um modelo de embeddings, o que permitiria a realização de testes mais amplos e a adaptação do sistema para diferentes cenários.

6 AGRADECIMENTOS

Agradecemos ao Instituto Federal do Tocantins (IFTO) e à Fundação de Amparo à Pesquisa do Tocantins (FAPT), pelo fomento e apoio concedido por meio do EDITAL Nº 58/2024 - IFTO/SEFAZ/FAPT/PIBIC, que tornou possível a realização desta pesquisa.

REFERÊNCIAS

DEMPORE, J.; MODUGU, K.; HESHAM, A.; RAMASAMY, L. **The impact of ChatGPT on higher education. Frontiers in Education**, 2023. Disponível em:https://doi.org/10.3389/feduc.2023.1206936. Acesso em: 29 set. 2024.

ILIEVA, G.; YANKOVA, T.; KLISAROVA-BELCHEVA, S.; DIMITROV, A.; BRATKOV, M.; ANGELOV, D. **Effects of Generative Chatbots in Higher Education.** Information, 2023. Disponível em: https://doi.org/10.3390/info14090492. Acesso em: 29 set. 2024.

MENDOZA, S.; SÁNCHEZ-ADAME, L.; URQUIZA-YLLESCAS, J.; GONZÁLEZ-BELTRÁN, B.; DECOUCHANT, D. **A Model to Develop Chatbots for Assisting the Teaching and Learning Process**. Sensors (Basel, Switzerland), 22, 2022. Disponível em: https://doi.org/10.3390/s22155532. Acesso em: 29 set. 2024.

PÉREZ, J.; DARADOUMIS, T.; PUIG, J. **Rediscovering the use of chatbots in education**: A systematic literature review. Computer Applications in Engineering Education, 2020. Disponível em: https://doi.org/10.1002/cae.22326. Acesso em: 29 set. 2024.

LEE, K.; JO, J.; KIM, J.; KANG, Y. Can chatbots help reduce the workload of administrative officers? Implementing and deploying FAQ chatbot service in a university. In: **INTERNATIONAL CONFERENCE ON HUMAN-COMPUTER INTERACTION**, 21., 2019, Orlando. Proceedings... Cham: Springer, 2019. Disponível em: https://doi.org/10.1007/978-3-030-23522-2_45. Acesso em: 22 abr. 2025.

LI, Y. **The potential application of ChatGPT in higher education management.** Lecture Notes in Education Psychology and Public Media, [S. l.], v. 25, p. 7439, 2023. Disponível em: https://doi.org/10.54254/2753-7048/25/20230750. Acesso em: 22 abr. 2025.