"Planeta Água: a cultura oceânica para enfrentar as mudanças climáticas no meu território"
SECT - ICE
XII Semana de Ciência e Tecnologia SECT ICE
20 a 23 de Outubro de 2025
Realização:

# ChatICE: Um Sistema para Auxiliar o Atendimento aos Discentes do ICE Baseado em RAG (Retrieval-Augmented Generation)

Evellim Michele Silva Martins[1,2](PQ), Carlos Daniel da Silva Ribeiro[1](PQ), Diogo Soares Moreira[1*](PQ), Ewerton Rodrigo Nunes Petillo[1](PQ), Luiz Gabriel Antunes Sena[1](PQ).

[1]Universidade Federal do Amazonas, Centro de Tecnologia da Informação e Comunicação (CTIC), Av. Rodrigo Otávio Jordão Ramos, 6200, Coroado I, 69080-900, Manaus AM, Brasil.

[2]Universidade Federal do Amazonas, Departamento de Matemática (DM - ICE), Av. Rodrigo Otávio Jordão Ramos, 6200, Coroado I, 69080-900, Manaus AM, Brasil.

* *evellim.martins@ufam.edu.br*

**Palavras-Chave:** Inteligência Artificial, chatbot, Retrieval Augmented Generation, RAG.

## Introdução

A crescente adoção de chatbots reflete a demanda por agilizar o acesso às informações em diversas áreas, desde o setor público até o privado[1]. Nesse contexto, os LLMs (Large Language Models) surgem como alternativa para interpretar nuances e aprender com interações, além de responder com agentes conversacionais que combinam precisão informacional e comunicação humanizada[2,3].

Por outro lado, o uso da IA e LLMs na educação pode ir além, aprimorando o acesso à informação e a facilitação da comunicação entre discentes e informações institucionais. Para isso, a técnica RAG (Retrieval Augmented Generation) permite potencializar os LLMs em domínios informacionais específicos, ao combinar a geração de conteúdo com a recuperação automática de informação de uma base de conhecimento externa, antes da formulação de respostas[2].

Assim, este trabalho propõe o desenvolvimento e avaliação de um chatbot inteligente, denominado **ChatICE**, que utiliza a técnica RAG e LLMs, direcionado às necessidades informacionais da comunidade discente do ICE-UFAM. O sistema alimenta-se de dados institucionais estruturados, como horários, ementas, informações docentes e calendários acadêmicos. O objetivo central é suprir a fragmentação informacional e oferecer um canal unificado para consultas do cotidiano acadêmico.

A validação proposta busca não apenas demonstrar a viabilidade técnica da aplicação, mas também avaliar sua potencial efetividade como ferramenta de suporte ao aluno.

## Metodologia

A arquitetura do **ChatICE** foi concebida seguindo o paradigma RAG, integrando de forma coesa os módulos de coleta de dados, recuperação de informação e geração de respostas. O sistema foi desenvolvido para operar como uma aplicação web robusta, conforme detalhado nos fluxos abaixo:

A base de conhecimento foi construída de forma automatizada, através das seguintes etapas:

- Web Scraping: Implementou-se um scraper utilizando Python, que extraiu informações estruturadas do site do ICE-UFAM. Foram coletados dados como ementas de disciplinas, grades curriculares, horários de aula, informações de docentes e dos cursos;
- Pré-processamento: Os dados extraídos foram limpos (remoção de tags HTML) e consolidados em documentos no formato markdown.
- Pós-processamento: O markdown gerado na etapa anterior foi, então, submetido para a API do Gemini para refatoração do markdown.

O núcleo do sistema de recuperação foi implementado com uma classe (RetrieveRelevantText), responsável por transformar o conhecimento textual em uma base vetorial consultável. O conteúdo textual é segmentado em trechos de tamanho de 1024 caracteres com sobreposição de 128 caracteres utilizando o LlamaIndex. Os chunks são convertidos em vetores numéricos (embedding) por meio do modelo BAAI/bge-m3, otimizado para a língua portuguesa, e indexado com a biblioteca FAISS, para buscas por similaridade.

Por fim, foi adotada a estratégia de recuperar inicialmente $3 \times k$ chunks com base em similaridade semântica, e então selecionar no máximo $max(1, floor(k / \delta)\, f)$ chunks por arquivo, onde k é o número de chunks desejados na resposta final, e δ é o número máximo de arquivos distintos permitidos na seleção.

A geração das respostas é feita, combinando o contexto recuperado nos passos anteriores com LLMs. O sistema foi projetado para operar tanto com modelos locais, via Ollama, quanto com o modelo Gemini da Google. Adicionalmente, um mecanismo de rate limiting foi implementado para garantir conformidade com as restrições de uso da API. O prompt enviado ao LLM foi elaborado, incluindo o contexto recuperado, a pergunta do usuário e instruções específicas (system prompt).

A visão geral do sistema pode ser observada na Fig. 1.

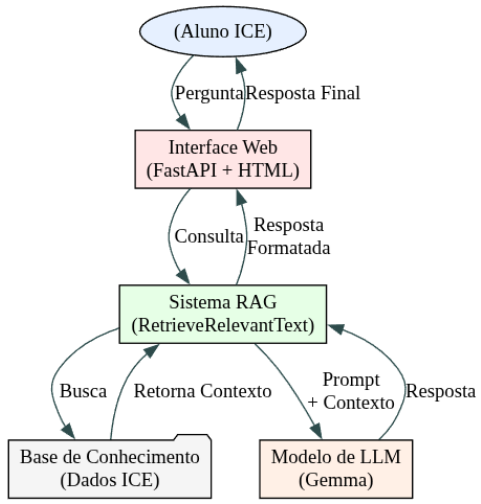


Figura 1. Diagrama de funcionamento do chatbot.

Para disponibilizar o chatbot, foi desenvolvida uma aplicação web com o framework FastAPI para gestão do chat e de sessões. No Frontend, a interface foi desenvolvida com HTML5, CSS3 e JavaScript, utilizando-se a biblioteca Jinja2 para renderização de interface. O sistema foi implantado em uma infraestrutura de nuvem, para visualização e testes.
A tabela 1 resume os parâmetros usados na construção do chatbot.

| Parâmetro | Valor |
|---|---|
| $k$ | 3 |
| $\delta$ | 7 |
| Tamanho Base Conhecimento | 20 arquivos .md |
| Tamanho médio por .md | 2000 caracteres |

## Resultados e Discussão

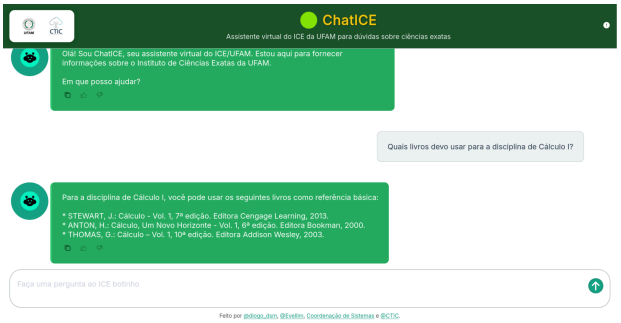


Figura 2. Exemplo 1 de pergunta no ChatICE.

Para testar o ChatICE foram realizadas perguntas dentro do contexto de necessidade do ICE, e analisadas as respostas com o intuito de verificar a qualidade dos resultados obtidos.
Nos exemplos apresentados nas Figuras 2 e 3, demonstram-se a realização das seguintes perguntas:

1. *Quais livros devo usar para a disciplina Cálculo I?*
2. *Quais são os pré-requisitos da disciplina de Métodos Numéricos I?*

Os resultados obtidos mostram a capacidade do ChatICE de extrair, corretamente, as informações obtidas dos arquivos presentes no site do ICE. Para as perguntas supracitadas, encontram-se as seguintes respostas:

1. Para a disciplina de Cálculo I, você pode usar os seguintes livros como referência básica:

   * STEWART, J.: Cálculo - Vol. 1, 7ª edição. Editora Cengage Learning, 2013.
   * ANTON, H.: Cálculo, Um Novo Horizonte - Vol. 1, 6ª edição. Editora Bookman, 2000.
   * THOMAS, G.: Cálculo – Vol. 1, 10ª edição. Editora Addison Wesley, 2003.

2. Para cursar Métodos Numéricos I, é necessário ter cursado Equações Diferenciais Ordinárias.


Figura 3. Exemplo 2 de pergunta no ChatICE.

## Conclusões

Neste trabalho, apresentamos o ChatICE, um bot que, dada a extração de páginas web do site do ICE usando um scraper, pode coletar informações do html da página ou de arquivos, armazenados no drive.
Ademais, as informações passam por refinamento através de IA para serem melhor reestruturadas. Contudo, o sistema depende da qualidade e da forma como as informações são disponibilizadas no site, e a busca frequentemente extrai conteúdos desconexos. Com isso, concluí-se que é necessário aprofundar a melhor maneira de estruturar esses conteúdos para o RAG, visando melhorar o desempenho do chat.

## Referências

[1] Cabral, J. P. F. (2025). Chatbot inteligente para acesso a regulamentos acadêmicos: um sistema de recuperação de informações baseado em RAG.

[2] Carlos, L. L. (2025). RAG acadêmico: facilitando o acesso à informação com LLMs e repositórios abertos de universidades.

[3] Santos, A., Santos, B. S., de Oliveira, R. P., & Silva, M. (2025, June). Lusi: um chatbot baseado em Modelos de Linguagem para auxiliar nos atendimentos ao publico em departamentos acadêmicos. In *Simpósio Brasileiro de Sistemas Colaborativos (SBSC)* (pp. 1-5). SBC.

[4] Pegano, B.; Caterino, M.; Filosa, R.; Giancola, C. Binding of harmine derivatives to dna: a spectroscopic investigation. *Molecules*, 22: 1831-1838. 2017.