

LIA: Um Chatbot para Consulta de Diretrizes de Inteligência Artificial em Universidades Brasileiras

Anna Myllenne Araújo dos Santos¹, Maria Jane de Queiroz¹, Heraldo Gonçalves Lima Junior¹, Danielle Rousy Dias Ricarte¹, Natasha Correia Queiroz Lino¹

¹Centro de Informática – Universidade Federal da Paraíba (UFPB)
CEP: 58058-600 – João Pessoa – PB – Brasil

{anna.myllenne,maria.jane.queiroz,heraldo.lima}@academico.ufpb.edu.br,
{danielle,natasha}@ci.ufpb.br

Abstract. *This work presents a chatbot based on Retrieval-Augmented Generation (RAG) to answer questions about regulations and guidelines on the use of artificial intelligence in Brazilian federal universities. The system employs a hybrid information retrieval approach, combining ChromaDB and BM25, along with the BAAI reranker model, in order to improve document retrieval accuracy. As a result, an accessible application was developed that provides natural language responses supported by reliable sources and demonstrates superior performance in the precision of retrieved information.*

Keywords: Generative AI; RAG; Regulations; Chatbot.

Resumo. *Este trabalho apresenta um chatbot baseado em Retrieval-Augmented Generation (RAG) para responder a perguntas sobre as regulamentações e guias do uso de inteligência artificial em universidades federais brasileiras. O sistema utiliza uma abordagem híbrida de recuperação de informação, combinando ChromaDB e BM25, juntamente com o modelo de reranqueamento BAAI, a fim de aumentar a acurácia na busca por documentos. Como resultado, foi desenvolvido um aplicativo acessível que fornece respostas em linguagem natural com suporte em fontes confiáveis e que apresenta resultados superiores na precisão da informação.*

Palavras-chave: IA Generativa; RAG; Regulamentação; Chatbot.

1. Introdução

Recentemente, a Inteligência Artificial (IA) tem sido impulsionada por tecnologias como *big data*, aprendizado de máquina e aprendizado profundo. Suas diversas aplicações em áreas como educação, saúde, finanças e transportes têm despertado atenção em escala global. Entre os avanços recentes, pode-se destacar a Inteligência Artificial Generativa (IAGen), que é caracterizada pela capacidade de modelos gerarem novos conteúdos como textos e imagens a partir de padrões identificados em grande volume de dados [Andifes, 2025].

Na área da educação, essa tecnologia oferece ferramentas promissoras para a inovação pedagógica, podendo ser usada para personalização do aprendizado [Ayeni et al., 2024], criação de materiais didáticos adaptativos [Machado et al., 2024], tutoria virtual [Alier et al., 2025; Falcão et al., 2020], análise da aprendizagem [Wahdiana et al., 2025; Vierhauser et al., 2024], entre outros.

A adoção acelerada da IA por professores e alunos tem contribuído para que as instituições ao redor do mundo promovam debates e formulem políticas para regulamentar o uso dessas tecnologias no cenário educacional [Moorhouse et al., 2023]. As instituições brasileiras também têm iniciado seu movimento regulatório, porém ainda de forma isolada, fragmentada e embrionária [Pires et al., 2025].

De acordo com Ghimire e Edwards (2024), as políticas de IA nas universidades têm avançado significativamente; porém, faz-se necessário que as instituições adaptem suas documentações para que as ideias não se tornem obsoletas diante da constante evolução da tecnologia, o que exige que sejam multimodais e flexíveis.

Todavia, essa natureza dinâmica das regulamentações introduz um novo desafio: garantir que os corpos discente, docente e administrativo permaneçam a par das atualizações que estão em constante transformação. De acordo com Rayzer e Santos (2026), a comunidade acadêmica tem dificuldade em acessar as informações devido a dispersão das normativas. Os autores também apontam que apenas o Conselho Universitário da Universidade do Estado de Santa Catarina, entre 2023 e 2025, definiu 214 novas resoluções, ressaltando que várias outras entidades internas à universidade, como conselhos, câmaras universitárias e centros de ensino também podem emitir suas próprias normativas [Rayzer & Santos, 2026]. O volume de documentos evidencia a necessidade de mecanismos que facilitem a intermediação entre as regulamentações vigentes e o uso efetivo da IA pela comunidade acadêmica.

Diante disto, abordagens baseadas em *Retrieval-Augmented Generation* (RAG, Geração Aumentada por Recuperação) mostram-se como uma possível solução. Essa técnica combina modelos de linguagem com mecanismos de recuperação de informação em bases documentais confiáveis, permitindo gerar respostas fundamentadas em fontes específicas e reduzindo problemas como a imprecisão factual ou a ausência de rastreabilidade [Lewis et al., 2020].

Ao integrar a geração de linguagem natural à consulta direta a documentos institucionais, os sistemas baseados em RAG possibilitam um acesso mais transparente, contextualizado e verificável às normas e recomendações organizacionais. Com isso, este trabalho tem como objetivo aplicar técnicas de RAG para criar um agente conversacional, capaz de responder, em linguagem natural clara, a questionamentos de professores e alunos sobre o que as universidades brasileiras abordam em suas respectivas diretrizes e recomendações de uso ético de IAGen nas instituições.

2. Referencial Teórico

Este referencial teórico organiza-se em três eixos: IAGen na Educação, Geração Aumentada por Recuperação (RAG) e regulamentação da IA nas Instituições de Ensino Superior. Esses temas permitem discutir as possibilidades de uso da IA no ensino, na aprendizagem e na avaliação, bem como os desafios éticos, acadêmicos e institucionais associados à sua adoção.

2.1. IAGen na Educação

O termo inteligência artificial foi primeiramente introduzido por John McCarthy em 1956 durante uma conferência em Dartmouth [Kaplan & Haenlein, 2019]. Décadas se

passaram e agora a tecnologia discutida pelo cientista, em específico a subseção da IAGen, ganhou novas dimensões com a popularização do *ChatGPT* em 2022, que vem promovendo rápida adoção social e educacional dessas tecnologias [Kasnezi et al., 2023].

A IAGen teve seu uso intensificado no ensino superior e o relatório da Organização para a Cooperação e Desenvolvimento Econômico (OCDE, 2026) evidencia essa crescente adoção. A síntese aponta um estudo francês de 2023, realizado com 4500 estudantes universitários, que indicou que aproximadamente 55% dos alunos já utilizam a tecnologia em seu cotidiano [Compilatio, 2023], e que essa parcela já aumentou para 82%, segundo Pascal et al. (2025). Um outro estudo realizado por Rong e Chun (2024) constatou que 86% dos alunos participantes da pesquisa haviam utilizado IA em seus estudos. Na mesma linha, o estudo realizado por Hüscher et al. (2025), com mais de 23.000 discentes alemães, relatou que 94% dos estudantes utilizaram IA em 2025, dos quais 65% a utilizavam diariamente ou semanalmente.

Diversos estudos recentes têm investigado os possíveis benefícios da integração da IA com o ensino superior e já é possível elencar alguns potenciais pedagógicos interessantes da tecnologia, como tutores virtuais [Al-Abri, 2025; Alier et al. 2025] personalização de aprendizagem [Murtaza et al. 2022], geração de *feedback* automático [Escalante et al. 2023], auxílio na escrita acadêmica [Alqahtani et al. 2023], geração de conteúdo, sistema de recomendação, predição de performance, entre outros usos benéficos [Castilho et al. 2025].

2.2. Geração Aumentada por Recuperação (Rag)

Agentes conversacionais baseados em *Large Language Models* (LLMs), ou Modelos de Linguagem de Grande Escala, têm revolucionado a forma como as informações podem ser acessadas hoje em dia, apresentando benefícios decisivos quando comparados com sistemas tradicionais [Klesel & Wittmann, 2025]. Porém, tais modelos contam com limitações estruturais relacionadas à factualidade e à sua capacidade de generalização, por serem limitados a dados paramétricos de seu treinamento [Bender et al., 2021] ou a informações públicas presentes na Internet [Brown et al., 2020], não tendo contexto específico para fornecer respostas assertivas em alguns cenários, como em casos específicos da medicina [Pal et al., 2023].

Diante desse impasse, surgiu a RAG, paradigma que serve como estratégia para amenizar tais limitações ao implementar a recuperação de informações externas dos dados de treinamento para os modelos de IAGen.

Diferentemente de abordagens clássicas, a RAG não necessita do retreinamento do modelo, pois sua estrutura, que consiste na adição de acesso à memória paramétrica e à não paramétrica, permite desprender-se da dependência das informações implicitamente codificadas nos pesos [Lewis et al., 2020; Gao et al., 2023]. Esta é uma arquitetura que combina um componente de recuperação (*retriever*), responsável por buscar documentos relevantes em uma base externa, com outro componente de geração (*generator*), tipicamente um modelo de linguagem pré-treinado, que produz a resposta final condicionada às evidências recuperadas [Lewis et al., 2020].

Existem várias vantagens para uma implementação baseada em RAG; entre elas, pode-se citar melhoria da precisão factual [Lewis et al., 2020], rastreabilidade de respostas [Lewis et al., 2020], ganhos no desempenho de tarefas de perguntas e respostas (*question-answering*), principalmente no domínio de sistemas de FAQ [Gao et al., 2023; Kleser & Wittmann, 2025]. É igualmente valioso em sistemas de resposta para contextos específicos [Kleser & Wittmann, 2025].

No entanto, a literatura também aponta várias limitações desta arquitetura, como a presença de alucinações, que não é totalmente evitada com o RAG (Gao et al., 2023; Li et al., 2025); a dependência dos documentos, que leva a respostas repetitivas e redundantes (Gao et al., 2023); elevados custos computacionais (Li et al., 2025); apoio multimodal limitado, uma vez que a maioria das aplicações se concentra no texto (Li et al., 2025); entre outras.

2.3. Estado Regulamentatório das Instituições de Ensino Superior

A incorporação da IAGen no sistema educacional tem sido acompanhada de esforços na elaboração de regulamentações e na estruturação de políticas de governança, tanto por universidades quanto por instituições terceirizadas.

No cenário nacional, é possível ver a movimentação orgânica das instituições para a normatização e governança com o surgimento de recomendações e políticas institucionais para padronizar práticas, mitigar riscos e orientar usos pedagógicos [Andifes, 2025], mas tal movimento ainda se encontra em estado embrionário [Pires et al., 2025]. De acordo com pesquisa realizada pela UNESCO em setembro de 2025, com a participação de 400 universidades de 90 países, enquanto os países da Europa e da América do Norte contam com aproximadamente 70% das instituições com políticas já instituídas ou em andamento, somente 45% das instituições da América Latina e do Caribe estão nessa fase de desenvolvimento [UNESCO, 2025].

Dentre as universidades nacionais que já dispõem de regulamentações consolidadas, observam-se diferentes interpretações e níveis de rigidez. Por exemplo, a Universidade Federal da Paraíba (UFPB) [UFPB, 2025] e a Universidade Federal do Ceará (UFC) estabelecem limitações claras sobre o que é permitido ou vedado nas instituições, estipulando sanções que vão desde não conseguir marcar a defesa da tese na UFC [UFC, 2025] até a cassação de diploma, a necessidade de adaptação de trabalhos e até a retração de artigos publicados pela comunidade acadêmica da UFPB.

Por outro lado, universidades como a Universidade Federal de Minas Gerais (UFMG) e a Universidade Federal da Bahia (UFBA) apresentam políticas mais flexíveis e não taxativas, publicando guias de bom uso e recomendações para orientar o corpo acadêmico sobre um uso mais consciente e ético da IA sem estipular sanções [UFMG, 2024; UFBA, 2025]. Tais distinções indicam que o estado regulatório brasileiro encontra-se em fase de consolidação.

3. Trabalhos Relacionados

É possível encontrar diversas soluções que aplicam o RAG para gerar respostas relevantes ao contexto universitário. Por exemplo, o trabalho de Amaral et al. (2025) desenvolveu um *chatbot* com RAG para auxiliar nas demandas da administração de

graduação da Universidade Federal de Santa Maria (UFSM), integrando seu *bot* ao Telegram e possibilitando que alunos tirassem dúvidas sobre assuntos como currículo, trabalho de conclusão de curso, estágio, entre outros que fossem abrangidos em documentos regulatórios.

Na mesma direção, Silva e Nascimento (2023) propuseram o *chatbot* Polímata, uma ferramenta de código aberto que responde às dúvidas frequentes feitas pelos alunos da Universidade Federal do Amapá (UNIFAP). Sob a mesma perspectiva, Nascimento et al. (2025) desenvolveram o “BichoBot”, uma solução que fornece informações importantes a alunos ingressantes na UFC, combinando técnicas de RAG com o *framework* Rasa.

Outras iniciativas que atacam o mesmo problema no contexto do ensino superior são o “UniChat” [Bordin et al., 2025] da UFSM; o trabalho desenvolvido por Rayzer e Santos (2025) dedicado às normativas da Universidade do Estado de Santa Catarina (UDESC); o “Sabiá” [Rodrigues et al., 2025] projetado para auxiliar a comunidade acadêmica da Universidade Tecnológica Federal do Paraná (UTFPR); o *chatbot* “Freire” desenvolvido por Mesquita et al. (2025) com enfoque na Universidade Federal do Ceará e o “IFBot”, que estuda a aplicação do RAG para resolver o problema no contexto da instituição dos autores, o Instituto Federal Fluminense [Souza et al., 2025].

A partir disso, o presente artigo se diferencia dos trabalhos apresentados neste tópico em duas dimensões principais, com relação ao escopo documental e à arquitetura de recuperação desenvolvida para o *chatbot*. Quanto ao escopo, enquanto os trabalhos anteriores tinham como base as normativas de uma única universidade, focando em temas como currículo, estágio e informações administrativas, o presente trabalho indexa documentos de diversas universidades espalhadas pelo território nacional, tendo como público-alvo estudantes, docentes e servidores de todas as instituições federais de ensino superior que já contam com regulamentação para o uso de IA.

Quanto à arquitetura, os trabalhos relacionados apresentam, predominantemente, estratégias de recuperação simples, baseadas principalmente em buscas semânticas por similaridade vetorial. Por outro lado, o *pipeline* implementado neste trabalho se inspira na arquitetura proposta por Sun et al. (2024), em que ocorre a combinação de dois mecanismos complementares de recuperação, o BM25 (*Best Match 25*) e o FAISS. Após isso, os dois retrievers são fundidos e os candidatos recuperados são reordenados antes de serem enviados ao modelo de LLM. Neste estudo, o FAISS é substituído pelo ChromaDB, devido ao seu suporte nativo para busca filtrada. Além disso, o modelo do reranqueador foi substituído pelo BAAI, que se adequa melhor ao português brasileiro.

Com isso, pode-se perceber que esta pesquisa diferencia-se do que já existe na literatura nacional e evidencia-se uma oportunidade de examinar a eficácia da aplicação de uma arquitetura desenvolvida para um contexto diferente do que foi previamente discutido.

4. Métodos e Materiais

Esta seção aborda como foi feito o *pipeline* do *chatbot* em detalhes, apresentando o processo de busca e organização da documentação que compõe o *dataset*.

4.1. Composição do *Dataset*

A presente pesquisa restringiu seu escopo a apenas instituições de ensino superior federais brasileiras. Essa escolha metodológica fundamenta-se na busca por maior rigor e homogeneidade, a fim de garantir maior reprodutibilidade do processo de coleta. Dessa forma, a delimitação proposta viabiliza a replicação da metodologia por pesquisadores futuros, uma vez que o número de instituições federais é limitado e seus portais são de acesso público, o que facilita o acesso às informações.

A construção do *corpus* documental que serviu para alimentar o *chatbot* foi operacionalizada em duas etapas: a primeira, uma busca automatizada via *web scraping* e a segunda, caracterizada pela curadoria manual dos documentos institucionais encontrados. A busca foi realizada nos portais oficiais de cada uma das instituições, o que permitiu a adoção de uma estratégia de rastreamento direto, recuperando documentos publicados até o dia 20 de março de 2026. Os termos de busca contemplaram expressões como “*regulamentação inteligência artificial*”, “*política de uso de IA*”, “*diretrizes inteligência artificial*”, “*guia de uso de IA*”, entre outros. Após essa etapa, iniciou-se a revisão manual para garantir a autenticidade, a relevância e a exclusão de arquivos duplicados.

Os critérios de inclusão adotados foram: (i) o documento deve ter sido formalmente produzido por uma universidade federal brasileira; (ii) o documento deve abordar o uso da IA no contexto acadêmico interno da instituição; (iii) o documento deve estar publicamente acessível, de forma integral, para garantir reprodutibilidade; (iv) o documento não deve limitar seu escopo a um curso específico. Após a aplicação desses critérios, o *dataset* final foi composto por documentos de 16 universidades, todos em formato PDF (*Portable Document Format*), conforme o Quadro 1.

Quadro 1. Síntese das universidades que compõem o dataset.

Universidade	Escopo	Natureza do documento	Região
UFBA (2025)	Ensino, pesquisa, extensão e administração	Guia	Nordeste
UFC (2025)	Pós-graduação	Diretriz	Nordeste
UFDPAr (2025)	Ensino, pesquisa, extensão e administração	Diretriz	Nordeste
UFF (2025)	Ensino, pesquisa, extensão e administração	Guia	Sudeste
UFG (2024)	Ensino, pesquisa, extensão e administração	Guia	Centro-Oeste
UFMA (2025)	Ensino, pesquisa, extensão e administração	Diretriz	Nordeste
UFMG (2024)	Ensino, pesquisa, extensão e administração	Recomendações	Sudeste
UFMS (2025)	Ensino, pesquisa, extensão e administração	Diretriz	Centro-Oeste
UFOP (2025)	Pós-graduação	Diretriz	Sudeste
UFPB (2025)	Ensino, pesquisa e extensão	Diretriz	Nordeste
UFRJ (2025)	Ensino, pesquisa, extensão e administração	Recomendações	Sudeste
UFRRJ (2025)	Ensino, pesquisa e extensão	Guia	Sudeste
UFU (2025)	Ensino, pesquisa, extensão e administração	Guia	Sudeste
UNIFAL-MG(2025)	Ensino, pesquisa e extensão	Diretriz	Sudeste
UNIFESP (2025)	Pós-graduação e pesquisa	Diretriz	Sudeste
UNIR (2026)	Pós-graduação	Diretriz	Norte

Os documentos foram classificados entre as categorias de guia, recomendações e diretrizes. Os guias apresentam natureza não taxativa, porém com elevado grau de formalidade, com discussões mais aprofundadas sobre os temas abordados e apresentação clara do que caracteriza o uso adequado. As recomendações constituem

documentos de menor densidade técnica, nos quais são abordados alguns pontos relevantes, mas a instituição não assume posicionamentos incisivos, servindo como uma apresentação da temática. Já as diretrizes estabelecem regras claras do que a universidade aceita e espera da comunidade, cujo descumprimento estão passíveis a punições institucionais.

4.2. Pipeline do Chatbot

O *pipeline* do LIA é dividido em duas fases principais: a primeira é a indexação dos documentos reunidos e a segunda é a geração de respostas com base nos trechos que o sistema de recuperação é capaz de retornar. O processo de indexação trata os documentos institucionais, organizando-os em estruturas de dados que permitem a recuperação eficiente, e é executado apenas uma vez. Por outro lado, a etapa de geração de respostas opera em tempo real a cada nova consulta do usuário, recuperando os segmentos mais relevantes à pergunta específica feita ao sistema. A figura 1 resume o processo de indexação que será explicitado.

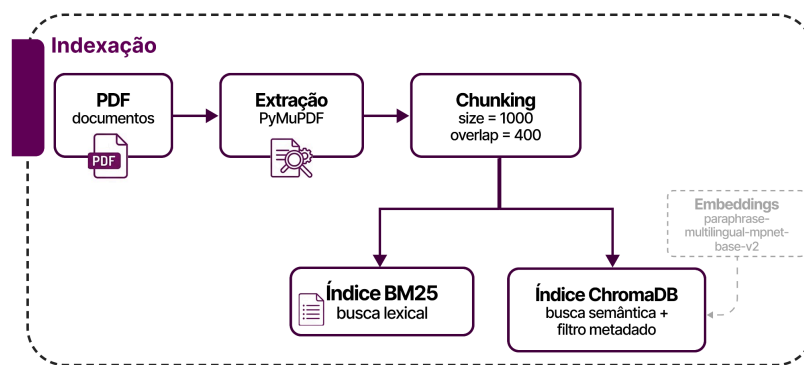


Figura 1. Detalhamento do pipeline do processo de Indexação.

A fase de indexação percorre três etapas sequenciais: extração de texto, segmentação e, por fim, indexação propriamente dita. Primeiramente, foi realizada a extração do conteúdo textual dos arquivos, em nível de página. Os parâmetros escolhidos foram 1000 para o tamanho máximo de caracteres por segmento (*chunk-size*) e 400 para o tamanho de sobreposição entre segmentos consecutivos (*chunk-overlap*). Esse tamanho de sobreposição tem como objetivo preservar a continuidade semântica entre os trechos adjacentes que poderia ser perdida pelo truncamento do texto.

Após esse processo, os segmentos passados para o ChromaDB passam a ser representados por um vetor denso de alta dimensionalidade. Esses vetores resultantes são indexados a uma coleção do ChromaDB [Huber & Troynikov, 2022], um banco de dados vetorial que armazena segmentos, associando cada um deles ao respectivo vetor e aos seus metadados. O ChromaDB realiza busca por similaridade de cosseno, medindo o ângulo entre o valor da consulta e os vetores armazenados, retornando os trechos que apresentam representações semânticas geometricamente mais próximas à pergunta formulada pelo usuário. Ele também permite a filtragem nativa por metadados, sendo essa capacidade utilizada para restringir a busca por segmentos a uma única universidade, a depender da pergunta do usuário, visando reunir segmentos de maior relevância.

Em paralelo, os mesmos segmentos são indexados através do algoritmo BM25, algoritmo probabilístico de recuperação de informação baseado no modelo TF-IDF, que fica responsável pela busca lexical dos documentos no *pipeline*. O BM25 atribui pontuações com base na frequência dos termos da consulta em cada trecho e na raridade desses termos no corpus, de forma geral [Robertson & Zaragoza, 2009]. Ele opera sobre a correspondência lexical entre os termos da consulta e os termos dos segmentos de forma exata, sendo uma técnica eficaz para recuperar trechos que contenham terminologia técnica ou expressões específicas. Feita essa indexação inicial, a próxima etapa foi a geração de respostas, processo esse sintetizado na Figura 2.

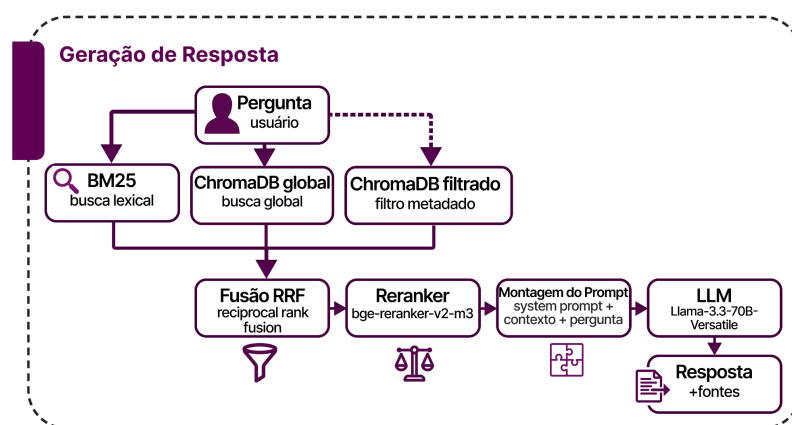


Figura 2. Detalhamento do processo de geração de resposta.

Para cada consulta realizada, são disparadas buscas por meio dos mecanismos de recuperação e os resultados são fundidos pelo algoritmo RRF (*Reciprocal Rank Fusion*), repassando-os para o modelo reranqueador, para que ocorra a reorganização. Após escolhidos os *k* segmentos mais importantes, esses são submetidos como contexto ao modelo de geração.

Para cada consulta feita, são identificadas as instituições mencionadas por meio de um dicionário de aliases, ou seja, seus nomes alternativos e, com base nessa detecção, o número de *retrievers* é decidido de forma dinâmica, sendo formado por uma busca BM25 e ChromaDB global como partes fixas e um número variável de buscas ChromaDB filtrado, sendo uma para cada universidade identificada, formando assim uma abordagem híbrida para recuperação que busca mitigar as limitações específicas de cada componente.

Os conjuntos de segmentos recuperados são combinados numa só lista ordenada pelo algoritmo RRF [Cormack et al., 2009]. Esse método de fusão atribui a cada segmento uma pontuação por meio do cálculo que se baseia nas posições que ele ocupa em cada lista individual, tendo como uma das suas principais vantagens em relação a modelos alternativos não necessitar que as pontuações absolutas dos diferentes recuperadores sejam comparáveis entre si, o que o torna interessante para combinar sistemas heterogêneos como o proposto neste trabalho [Sun et al., 2024].

Sob esse conjunto consolidado na etapa de fusão, é aplicado o estágio de reranqueamento neural utilizando o modelo *BAAI/bge-reranker-v2-m3*. O contexto final

conta com os top k segmentos relevantes reranqueados, sendo esse k um valor variado a depender do tipo de pergunta feita pelo usuário.

Todos os K segmentos são estruturados de forma lógica para que seja identificada a universidade a que pertencem e concatenados com a consulta feita pelo usuário. Em seguida, o *prompt* é então submetido ao modelo de linguagem (*Llama 3.3 70B Versatile*). O modelo opera sob um *system prompt* que lhe impõe limitações de comportamento.

5. Validação

Para validação do *chatbot*, foi utilizada uma estrutura de avaliação proposta por Gao et al. (2024), que trata da validação da qualidade da recuperação e da robustez da geração como partes distintas. Dessa forma, avaliou-se o processo de recuperação de segmentos de forma independente, usando as métricas P@K (*Precision at K*), R@K (*Recall at K*) e MRR (*Mean Reciprocal Rank*), provendo assim uma avaliação ponta a ponta por meio das métricas F1, *Precision* e *Recall*. As métricas P@K, R@K e MRR foram escolhidas para a avaliação isolada do componente de recuperação por serem amplamente utilizadas na literatura, além de ser importante compreender não apenas se os segmentos essenciais são recuperados, mas também qual a posição no *ranking* e em que quantidade eles aparecem [Liu, 2009].

O P@K mede qual a proporção de K trechos retornados pelo recuperador são realmente relevantes. O R@K é responsável por medir a proporção de *chunks* recuperados que estão dentro do total de *chunks* relevantes existentes no gabarito. Por fim, o MRR é a métrica que avalia a posição dos *chunks* relevantes em relação à lista dos que foram retornados e determina a proporção de segmentos realmente importantes entre as primeiras posições.

Para a avaliação ponta a ponta, seguiu-se o protocolo de avaliação utilizado por Rajpurkar et al. (2016). Esse conjunto de métricas foi escolhido por já ter sido validado no trabalho de Sun et al. (2024). Para isso, as respostas são normalizadas e comparadas *token a token* com a referência. O *Precision* mede quantos *tokens* gerados estão também presentes na referência, e o *Recall* mensura quantos *tokens* da referência estão presentes na resposta gerada, sendo o F1 uma junção dos dois em uma média harmônica.

Para obter os resultados em cada uma dessas métricas, foi construído um *dataset* composto por 90 perguntas elaboradas e anotadas manualmente com base nos documentos regulatórios indexados no *chatbot*. No conjunto, cada uma das 16 universidades recebeu cinco questões específicas, complementadas com 10 perguntas comparativas, em que cada instituição aparece ao menos uma vez. Esse conjunto, então, foi submetido ao *pipeline* fragmentado para que fosse possível identificar a importância de cada componente que constitui o *chatbot* e, após, ao *pipeline* completo para ser avaliada em comparação com um modelo de geração de texto sem RAG.

6. Resultados

A implementação do LIA se deu por meio do Streamlit, fornecendo um site acessível em que as consultas são respondidas com base nos documentos indexados e são fornecidas as fontes consultadas. A Figura 3 apresenta um exemplo de saída.

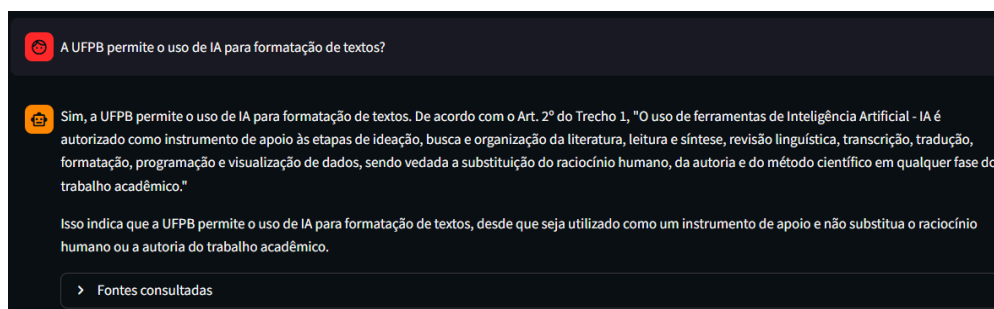


Figura 3. Exemplo de saída.

O sistema abrange 16 universidades, implementando um *framework* de recuperação com reranqueamento, um dos diferenciais deste estudo, conforme o Quadro 2, que compara a solução aqui proposta com os demais estudos analisados.

Quadro 2. Comparativo entre ferramentas que aplicam RAG.

Trabalhos	Escopo	Recuperação	Reranqueamento
LIA (este estudo)	16 universidades	ChromaDB + BM25	BAAI
Amaral et al.	UFMS	FAISS	ColBERTv2
Polimita	UNIFAP	Não declarado	-
BichoBot	UFC	Qdrant	-
Unichat	UFSM	PGVector	-
Sabiá	UTFPR	ChromaDB	-
Freire	UFC	ChromaDB	-
IFBot	IFF	FAISS	-
Rayzer e Santos	UDESC	FAISS	-

A avaliação, cujos resultados são apresentados na Tabela 1, indica que a melhor configuração é a fusão híbrida com junção do *reranker*, que obteve um desempenho notoriamente superior às outras. O alto valor na métrica R@K aponta que o *chatbot* recuperou uma boa parcela dos *chunks* considerados importantes. Seu desempenho de 69% na métrica MRR demonstra que os segmentos são classificados relativamente bem no seu nível de importância pelo *pipeline*, mas o seu baixo valor de 0.215 presente na métrica P@K aponta oportunidades de refinamento do *pipeline*, que recupera *chunks* que não são necessários para a formulação da resposta segundo o gabarito formulado.

Tabela 1. Resultados com variação do pipeline.

Configuração do pipeline	P@K	R@K	MRR
Somente BM25	0.078	0.339	0.252
Somente ChromaDB global	0.058	0.210	0.173
Somente ChromaDB filtrado	0.161	0.538	0.435
BM25 + ChromaDB global + ChromaDB filtrado	0.128	0.478	0.405
BM25 + Reranker	0.101	0.428	0.436
ChromaDB global + Reranker	0.070	0.237	0.271
BM25 + ChromaDB global + ChromaDB filtrado + Reranker	0.215	0.750	0.692

Logo atrás da arquitetura completa, a configuração que apresentou melhores resultados foi o ChromaDB filtrado, que superou até mesmo os resultados apresentados pela híbrida e pelos componentes combinados com o *reranker*. Tal resultado pode ser explicado pela limitação da busca somente entre trechos da universidade presentes na pergunta do usuário, o que diminui o ruído. No geral, o reranqueador apresenta impacto

positivo em todos os arranjos, mas não teve melhoras significativas ao ser combinado com o Chroma global, o que evidencia a importância de um processo de recuperação eficiente antes de passar para a etapa de reordenação de relevância dos segmentos.

O mesmo conjunto de questões anotadas foi novamente submetido ao *pipeline*, dessa vez completo, para comparar a eficiência da arquitetura com uma resposta por LLM pura, tendo sido utilizado o modelo *llama-3.1-8b-instant* para realizar os testes. Os resultados são apresentados na Tabela 2.

Tabela 2. Resultados RAG comparados com LLM puro.

Configuração	Precision	Recall	F1
Pipeline completo RAG	0.574	0.295	0.367
LLM pura	0.223	0.142	0.157
Δ (RAG-LLM)	0.352	0.153	0.210

O *pipeline* completo apresenta resultados melhores em todas as métricas, indicando que a arquitetura RAG auxilia o modelo a escrever respostas com maior adesão aos documentos indexados; porém, o resultado baixo em *Recall* pode indicar uma falha na construção do sistema do *chatbot*. Tal valor pode ter sido resultado da natureza da aplicação, que trabalha com documentação de linguagem técnica e específica ao contexto, e da natureza da métrica, que não considera paráfrases como corretas e pode ter prejudicado os resultados.

Quando comparado com os resultados obtidos por Sun et al. (2024), é possível perceber algumas similaridades. Ao realizar os testes com seu *pipeline* completo, os autores obtiveram um valor de F1 igual a 42.21%, com o resultado da métrica *Precision* retornando 47.29%. Esses valores se referem ao *pipeline* híbrido com *reranker* adicionado com o passo de *few-shot learning*, processo de oferecer ao modelo de geração exemplos de respostas construídas com base nos trechos, o que não é realizado no *chatbot* LIA, que utilizou *system prompt*.

Esse passo, que foi substituído no presente *pipeline*, se mostrou determinante nos resultados derivados dos experimentos de Sun et al. (2024). A configuração sem esse passo e sem o reranqueador apresentou valores de F1 iguais a 19.75%, e mesmo a arquitetura RAG completa obteve resultados inferiores ao que foi encontrado ao ser adicionado o passo de *few-shot learning*. Apesar das diferenças, as comparações permitem situar o desempenho do LIA em relação ao intervalo encontrado no trabalho de Sun et al. (2024), tomado como base para o desenvolvimento desta arquitetura.

7. Conclusão

Este estudo apresenta o desenvolvimento de um *chatbot* que utiliza RAG no seu *pipeline* para a criação de um sistema de perguntas e respostas voltado a responder questões da comunidade acadêmica das universidades federais brasileiras sobre as regulamentações do uso de IAGen. O produto final conta com uma estrutura híbrida, mesclando o ChromaDB e o BM25 para atingir uma maior acurácia na recuperação dos documentos. Ao fim, o trabalho fornece um aplicativo acessível ao público em geral que responde perguntas em linguagem humana provendo fontes que embasam as suas respostas.

Embora os resultados obtidos demonstrem o potencial da arquitetura proposta, este estudo apresenta algumas limitações que devem ser consideradas na interpretação dos achados. Em primeiro lugar, o *chatbot* foi desenvolvido com base em um conjunto documental restrito, abrangendo apenas documentos institucionais de 16 universidades. Além disso, a base de conhecimento é estática, não incorporando mecanismos automáticos de atualização, o que pode levar à obsolescência das informações à medida que as instituições modificam seus regulamentos e documentos oficiais.

Outra limitação refere-se ao processo de construção do conjunto de referência (gabarito), realizado por um único avaliador. Embora esse procedimento tenha permitido a condução dos experimentos, a ausência de múltiplos anotadores impossibilitou o cálculo de métricas de concordância interavaliadores (*inter-annotator agreement*), limitando a análise da consistência das anotações utilizadas na avaliação.

Também deve ser destacado que os hiperparâmetros empregados no pipeline, como o tamanho dos segmentos, da sobreposição e dos K escolhidos, foram definidos com base em recomendações da literatura, sem a realização de uma busca sistemática. Dessa forma, configurações alternativas podem produzir resultados distintos e potencialmente superiores. Além disso, os experimentos foram conduzidos utilizando um único LLM. Consequentemente, os resultados observados refletem o comportamento desse modelo específico, não sendo possível generalizar os achados.

Por fim, embora a avaliação experimental tenha demonstrado resultados promissores por meio de métricas objetivas, não foi realizada uma avaliação com usuários finais. Considerando que a proposta consiste em uma ferramenta interativa de apoio à comunidade acadêmica, estudos envolvendo o público alvo poderão fornecer evidências complementares sobre aspectos como usabilidade, utilidade percebida, satisfação e efetividade em cenários reais de utilização.

Como trabalhos futuros, recomenda-se ampliar o conjunto de universidades contempladas, desenvolvendo mecanismos automatizados de aquisição e atualização contínua da base documental; investigar estratégias de otimização dos hiperparâmetros do pipeline; comparar o desempenho da arquitetura utilizando diferentes modelos de linguagem; incorporar múltiplos especialistas no processo de anotação; e realizar estudos com usuários reais, complementados por métricas consolidadas de avaliação de sistemas RAG, como RAGAS e BERTScore, a fim de obter uma compreensão mais abrangente da qualidade, robustez e usabilidade da solução proposta.

Disponibilidade de Artefatos

Tanto o código-fonte do site quanto o pipeline RAG estão disponíveis no repositório: <https://github.com/wqffies/LIA-SBIE>

Declaração de Uso de IA

Foram utilizadas as seguintes ferramentas de IA Generativa neste artigo: ChatGPT (modelo GPT 5.3) como auxílio na revisão ortográfica e formatação das referências com base no padrão estabelecido pela SBC; Claude (modelo Sonnet 4.6) como apoio na tradução do resumo para inglês.

Referências

- Al-Abri, A. (2025). Exploring ChatGPT as a virtual tutor: A multi-dimensional analysis of large language models in academic support. *Education and Information Technologies*, v. 30, n. 12, p. 17447–17482.
- Alier, M., Pereira, J., García-Peñalvo, F. J., Casañ, M. J. and Cabré, J. (2025). LAMB: An open-source software framework to create artificial intelligence assistants deployed and integrated into learning management systems. *Computer Standards & Interfaces*, v. 92, p. 103940.
- Alqahtani, T. et al. (2023). The emergent role of artificial intelligence, natural learning processing, and large language models in higher education and research. *Research in Social and Administrative Pharmacy*, v. 19, n. 8, p. 1236–1242. DOI: 10.1016/j.sapharm.2023.05.016.
- Andifes, Associação Nacional dos Dirigentes das Instituições Federais de Ensino Superior (2025). Diretrizes para o uso responsável da inteligência artificial nas universidades federais brasileiras. Brasília, DF.
- Amaral, V. M. R., Espinosa, L., Rigo, P. D. and Lunardi, G. M. (2025). A RAG-powered academic chatbot with ontology-driven factual verification. In: *Anais do Workshop LAGO*. Porto Alegre: Sociedade Brasileira de Computação.
- Ayeni, O. O., Nancy Mohd Al Hamad, Onyebuchi Nneamaka Chisom, Blessing Osawaru, and Ololade Elizabeth Adewusi (2024). AI in education: A review of personalized learning and educational technology. *GSC Advanced Research and Reviews*, v. 18, n. 2, p. 261– 271.
- Bender, E. M., Gebru, T., McMillan-Major, A. and Shmitchell, S. (2021). 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)*, p. 610–623. DOI: 10.1145/3442188.3445922.
- Bordin, A. C. B., Lunardi, G. M., Piveta, E. K. and Emmendorfer, L. (2025). UniChat: Arquitetura e avaliação de um agente RAG de baixo custo para o contexto universitário. In: *Anais do Simpósio Brasileiro de Banco de Dados (SBBD)*, p. 886–892. Porto Alegre: Sociedade Brasileira de Computação. DOI: 10.5753/sbbd.2025.247790.
- Brown, T. B. et al. (2020). Language models are few-shot learners. *arXiv preprint*, p. arXiv:2005.14165. DOI: 10.48550/arXiv.2005.14165.
- Castilho, G. U., Herrera, V. A. and Rodriguez, C. L. (2024). Explorando o potencial da IA generativa em ambientes virtuais de aprendizagem: uma revisão sistemática. In: *Anais do Simpósio Brasileiro de Informática na Educação (SBIE)*. Porto Alegre: Sociedade Brasileira de Computação.
- Compilatio (2023), L'IA dans l'enseignement : résultats détaillés d'une enquête où étudiants et enseignants confrontent leurs regards. Disponível em: <https://www.compilatio.net/blog/enquete-ia-enseignement-2023#enquete>. Acesso em: 03 fev 2026.

- Cormack, G. V., Clarke, C. L. A. and Buettcher, S. (2009). Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In: *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '09)*, p. 758–759. DOI: 10.1145/1571941.1572114.
- Escalante, J., Pack, A. and Barrett, A. (2023). AI-generated feedback on writing: insights into efficacy and ENL student preference. *International Journal of Educational Technology in Higher Education*, v. 20, p. 57.
- Falcão, T. P., Arêdes, V., Souza, S., et al. (2020). Tutoria: uma Plataforma para Apoiar Boas Práticas de Feedback no Processo de Ensino e Aprendizagem. In Congresso Brasileiro de Informática na Educação (CBIE). SBC.
- Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, M. and Wang, H. (2024). Retrieval-augmented generation for large language models: A survey. arXiv preprint, p. arXiv:2312.10997. DOI: 10.48550/arXiv.2312.10997.
- Ghimire, A. and Edwards, J. (2024). Generative AI adoption in classroom in context of technology acceptance model (TAM) and the innovation diffusion theory (IDT). *arXiv preprint*, p. arXiv:2406.15360. DOI: 10.48550/arXiv.2406.15360.
- Huber, J. and Troynikov, A. (2022). Chroma: Open-source embedding database for AI applications. Disponível em: <https://github.com/chroma-core/chroma>. Acesso em: 10 mar. 2026.
- Hüsch, M., N. Horstmann and A. Breiter (2025), Künstliche Intelligenz im Studium – die Sicht von Studierenden im Wintersemester 2024/25. Disponível em: <https://hochschuldaten.ch.de/kuenstliche-intelligenz-im-studium-die-sicht-von-studierendenim-wintersemester-2024-25/>. Acesso em: 03 fev 2026.
- Kaplan, A. and Haenlein, M. (2019). Siri, Siri, in my hand: Who’s the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence. *Business Horizons*, v. 62, n. 1, p. 15–25. DOI: 10.1016/j.bushor.2018.08.004.
- Kasneci, E. et al. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, v. 103, p. 102274.
- Klesel, M. and Wittmann, H. F. (2025). Retrieval-augmented generation (RAG). *Business & Information Systems Engineering*, v. 67, n. 4, p. 551–561. DOI: 10.1007/s12599-025-00945-3.
- Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., and Kiela, D. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Advances in Neural Information Processing Systems (NeurIPS 2020)*, v. 33, p. 9459–9474. arXiv:2005.11401.
- Liu, T.-Y. (2009). Learning to rank for information retrieval. *Foundations and Trends in Information Retrieval*, v. 3, n. 3, p. 225–331. DOI: 10.1561/15000000016.
- Li, Z., Wang, Z., Wang, W., Hung, K., Xie, H. and Wang, F. L. (2025). Retrieval-augmented generation for educational application: A systematic survey.

- Computers and Education: Artificial Intelligence*, v. 8, p. 100417. DOI: 10.1016/j.caeai.2025.100417.
- Nascimento, S. O., Gonçalves, E., Oliveira, M. A. and Barros, E. F. (2025). BichoBot – A chatbot to assist in the reception of incoming university students. In: *Anais do Encontro Nacional de Inteligência Artificial e Computacional (ENIAC)*. Porto Alegre: Sociedade Brasileira de Computação. DOI: 10.5753/eniac.2025.13665.
- Machado, R., Nascimento, M., Bomfim, M., et al. (2024). Homero: Uma Plataforma de EaD com Inteligência Artificial para Aprendizagem Adaptativa e Personalizada. In Congresso Brasileiro de Informática na Educação (CBIE). SBC.
- Mesquita, A. W. et al. (2025). Hello, Freire! How can you help me? A smart chatbot to enhance access to academic opportunities and institutional matters. In: *Anais de evento da SBC*. Porto Alegre: Sociedade Brasileira de Computação.
- Moorhouse, B. L., Yeo, M. A., and Wan, Y. (2023). Generative AI tools and the evolving landscape of higher education: Policy considerations and pedagogical implications. *Computers and Education: Artificial Intelligence*, v. 4, 100152. DOI: 10.1016/j.caeai.2023.100152.
- Murtaza, M., Ahmed, Y., Shamsi, J. A., Sherwani, F. and Usman, M. (2022). AI-based personalized e-learning systems: Issues, challenges, and solutions. *IEEE Access*, v. 10, p. 81323–81342.
- OECD (2026). OECD Digital Education Outlook 2026: Exploring Effective Uses of Generative AI in Education. *Paris: OECD Publishing*. Disponível em: https://www.oecd.org/content/dam/oecd/en/publications/reports/2026/01/oecd-digital-education-outlook-2026_940e0dd8/062a7394-en.pdf. Acesso em: 03 fev 2026.
- Pal, A., Umapathi, L. K. and Sankarasubbu, M. (2023). Med-HALT: Medical domain hallucination test for large language models. *arXiv preprint*, p. arXiv:2307.15343. DOI: 10.48550/arXiv.2307.15343.
- Pascal, F., F. Taddéi, M. de Falco and É.-P. Gallié (2025), IA et Enseignement Supérieur: Formation, Structuration et Appropriation par la Société, Ministère chargé de l’Enseignement Supérieur et de la Recherche. Disponível em: <https://mission-ia-sup.forge.apps.education.fr/#/>. Acesso em: 26 dez 2025
- Pires, R., Moro, F. and Frigo, L. B. (2025). Educação e regulação da inteligência artificial nas universidades públicas brasileiras. *Cuaderno Activa*, v. 1, n. 1. DOI: 10.53995/20278101.2042.
- Rayzer, L. P. B. and Santos, F. (2026). Explorando o uso de LLMs com RAG para simplificar o acesso às normativas da Universidade do Estado de Santa Catarina. *Regae: Revista de Gestão e Avaliação Educacional*, v. 15, n. 24. DOI: 10.5902/2318133895063.
- Robertson, S. E. and Zaragoza, H. (2009). The probabilistic relevance framework: BM25 and beyond. *arXiv preprint*, p. arXiv:0904.3088.
- Rodrigues, G. B., Beal, F., Marcon, M., Souza, A. C. C., Ortoncelli, A. R., Souza, F. C. M. and Silva, R. A. (2025). Sabiá: Um chatbot de inteligência artificial generativa

- para suporte no dia a dia do ensino superior. *arXiv preprint*, p. arXiv:2511.10787. DOI: 10.48550/arXiv.2511.10787.
- Rong, H. and C. Chun (2024), Digital Education Council Global AI Student Survey 2024, Digital Education Council. Disponível em: <https://www.digitaleducationcouncil.com/post/digital-education-council-global-ai-student-survey-2024>. Acesso em: 03 fev 2026.
- Silva, L. and Nascimento, XX. (2023). Chatbot polímata: Um chatbot de código aberto para apoio acadêmico na unifap. Trabalho de conclusão de curso (bacharelado em computação), Universidade Federal do Amapá, Macapá, Brasil.
- Souza, W. F. V., Nogueira, I. L. and Souza, A. V. (2025). Assistente virtual inteligente para instituições educacionais: aplicação da técnica RAG no IFBot. In: *Anais do Workshop sobre as Implicações da Computação na Sociedade (WICS)*, p. 217–224. Porto Alegre: Sociedade Brasileira de Computação. DOI: 10.5753/wics.2025.7986.
- Sun, H., Wang, Y. and Zhang, S. (2024). Retrieval-augmented generation for domain-specific question answering: A case study on Pittsburgh and CMU. *arXiv preprint*, p. arXiv:2411.13691. DOI: 10.48550/arXiv.2411.13691.
- UNESCO. (2025) Two-thirds of higher education institutions have or are developing guidance on AI use. Disponível em: <https://www.unesco.org/en/articles/unesco-survey-two-thirds-higher-education-institutions-have-or-are-developing-guidance-ai-use>. Acesso em: 1 mar. 2026.
- Universidade Federal da Bahia (2025). Guia para uso ético e responsável da inteligência artificial generativa na Universidade Federal da Bahia. Salvador, BA.
- Universidade Federal do Ceará (2025). Portaria nº 39/PRPPG/UFC, de 01 de outubro de 2025. Dispõe sobre a obrigatoriedade de submissão de trabalhos acadêmicos à ferramenta de verificação de similaridade e regulamenta o uso de inteligência artificial. Fortaleza, CE.
- Universidade Federal de Minas Gerais (2024). Recomendações para o uso de ferramentas de inteligência artificial nas atividades acadêmicas. Belo Horizonte, MG.
- Universidade Federal da Paraíba (2025). Resolução nº 57/2025. Revoga a Resolução nº 53/2019 e institui a política de integridade acadêmica e científica no âmbito da Universidade Federal da Paraíba. João Pessoa, PB.
- Vierhauser, M., Sauerwein, C., Antensteiner, T., & Hatmanstorfer, S. (2024). Learning Analytics Support in Higher-Education: Towards a Multi-Level Shared Learning Analytics Framework. In *Proceedings of the 16th International Conference on Computer Supported Education*.
- Wahdiana, D., Ekohariadi, E., & Suhartini, R. (2025). Clustering Cadet Training Performance Using K-Means and Ward's Method Evidence from OTMon Maritime Monitoring System. *Journal of Applied Data Sciences*, 6(4), 2741-2757.