



**INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DO
TOCANTINS
CAMPUS ARAGUATINS
CURSO SUPERIOR DE LICENCIATURA EM COMPUTAÇÃO**

FÁBIO EMANUELL ABREU CARDOSO

**DESENVOLVIMENTO E AVALIAÇÃO DE UM AGENTE CONVERSACIONAL
PARA A GESTÃO ACADÊMICA DO IFTO UTILIZANDO LLMS COM RAG, BUSCA
HÍBRIDA E RERANKING**

ARAGUATINS
2025

FÁBIO EMANUELL ABREU CARDOSO

**DESENVOLVIMENTO E AVALIAÇÃO DE UM AGENTE CONVERSACIONAL
PARA A GESTÃO ACADÊMICA DO IFTO UTILIZANDO LLMS COM RAG, BUSCA
HÍBRIDA E RERANKING**

Trabalho de Conclusão de Curso apresentado à
Coordenação do Curso de Licenciatura em Computação
da Unidade Campus Araguatins, do Instituto Federal do
Tocantins, como exigência à obtenção do título de
Licenciado em Computação.

Orientador: Dr. Ramasio Ferreira de Melo

ARAGUATINS
2025

Dados Internacionais de Catalogação na Publicação (CIP)
Bibliotecas do Instituto Federal do Tocantins

C268d Cardoso, Fábio Emanuell Abreu
 Desenvolvimento e avaliação de um agente conversacional para a gestão acadêmica do IFTO utilizando LLMs com RAG, busca híbrida e Reranking / Fábio Emanuell Abreu Cardoso. – Araguatins, TO, 2025.
 35 p. : il. color.

 Trabalho de Conclusão de Curso (Licenciatura em Computação) – Instituto Federal de Educação, Ciência e Tecnologia do Tocantins, Campus Araguatins, Araguatins, TO, 2025.

 Orientador: Dr. Ramasio Ferreira de Melo

 1. Inteligência artificial. 2. Gestão acadêmica. 3. RAG híbrido com Reranking. I. Melo, Ramasio Ferreira de. II. Título.

CDD 004

A reprodução total ou parcial, de qualquer forma ou por qualquer meio, deste documento é autorizada para fins de estudo e pesquisa, desde que citada a fonte.
Elaborado pelo sistema de geração automática de ficha catalográfica do IFTO com os dados fornecidos pelo(a) autor(a).



Ministério da Educação
Secretaria de Educação Profissional e Tecnológica
Instituto Federal de Educação, Ciência e Tecnologia do Tocantins
Campus Araguatins

FOLHA DE APROVAÇÃO

TÍTULO DO TRABALHO: DESENVOLVIMENTO E AVALIAÇÃO DE UM AGENTE CONVERSACIONAL PARA A GESTÃO ACADÊMICA DO IFTO UTILIZANDO LLMS COM RAG, BUSCA HÍBRIDA E RERANKING

AUTOR: Fábio Emanuell Abreu Cardoso

ORIENTADOR: Prof. Dr. :Ramasio Ferreira de Melo

Trabalho de Conclusão de Curso apresentado ao Instituto Federal de Educação, Ciência e Tecnologia do Tocantins, *Campus Araguatins*, como parte das exigências para a conclusão do Curso de Licenciatura em Computação.

Aprovado em 03 de dezembro de 2025.

Prof. Dr. :Ramasio Ferreira de Melo
Presidente da Banca

Prof. Esp. Alan Gomes Silva
Primeiro avaliador

Prof. M.e Rogério Pereira de Sousa
Segundo avaliador



Documento assinado eletronicamente por **Ramasio Ferreira de Melo, Coordenador**, em 03/12/2025, às 12:05, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do [Decreto nº 8.539, de 8 de outubro de 2015](#).



Documento assinado eletronicamente por **ALAN GOMES SILVA, Servidor**, em 22/12/2025, às 09:31, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do [Decreto nº 8.539, de 8 de outubro de 2015](#).



Documento assinado eletronicamente por **Rogério Pereira de Sousa, Servidor**, em 25/02/2026, às 14:12, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do [Decreto nº 8.539, de 8 de outubro de 2015](#).



A autenticidade deste documento
pode ser conferida no site
[http://sei.iftto.edu.br/sei/controlador_externo.php?](http://sei.iftto.edu.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0)
[acao=documento_conferir&id_orgao_acesso_externo=0](http://sei.iftto.edu.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0), informando o
código verificador **2997164** e o código CRC **130B46EC**.

Referência: Processo nº 23233.024460/2025-51
2997164

SEI nº

Dedico este trabalho a mim mesmo, pela coragem de continuar avançando, mesmo quando o caminho pareceu difícil demais. Por ter seguido em frente com foco, força e fé, mesmo nos dias em que duvidei de mim. Dedico a este meu esforço silencioso, às batalhas internas vencidas e à perseverança que me trouxe até aqui. Dedico esta conquista a todos que caminharam ao meu lado nesta jornada. Cada gesto de cuidado, cada palavra de incentivo e cada presença sincera foram essenciais para transformar esforço em vitória e sonho em realidade.

AGRADECIMENTOS

Primeiramente, agradeço ao nosso Senhor Jesus Cristo, fonte de força, amparo e esperança. Sem Ele, nada disso seria possível. Foi Sua presença que iluminou meus passos, sustentou meus dias mais difíceis e me permitiu chegar até aqui com fé e determinação.

Agradeço de forma profundamente especial à minha família. À minha mãe, Jeane Abreu Saraiva Cardoso, pelo apoio imensurável, por cada esforço dedicado, por nunca medir sacrifícios para me ver avançar e, sobretudo, por seu amor incondicional, que me fortalece e me inspira a nunca desistir dos meus sonhos. Ao meu pai, Antonio Fábio Cardoso Pereira, por todo apoio, cuidado e por me ajudar a me tornar um homem íntegro, responsável e honrado, valores que levarei por toda a vida.

Aos meus irmãos, Gabriel Cardoso e Wallacy Hiorran, agradeço pelo carinho constante, pelo companheirismo e por serem parte essencial da minha história. Obrigado pelas conversas, pelas risadas, pelo apoio silencioso e pela presença que torna qualquer caminhada mais leve. À minha avó Maria da Conceição, agradeço por vibrar com cada passo meu, por suas orações cheias de amor e pela sua ternura encorajadora, que sempre me impulsionou a seguir adiante com confiança e esperança.

Aos meus amigos, levo comigo uma gratidão eterna. Agradeço, em especial, ao meu grande amigo Nicolas Heitor, companheiro de quatro anos de estudos intensos e de uma amizade genuína que nasceu ainda na infância. Juntos, atravessamos dias longos, noites cansativas e madrugadas silenciosas dedicadas ao aprendizado, aos desafios e ao nosso crescimento pessoal e acadêmico. Cada memória vivida ao seu lado é um tesouro afetivo que guardarei para sempre no coração. Obrigado por sempre estar ao meu lado, me apoiando, me auxiliando e construindo comigo uma verdadeira vida de irmandade. Meu sincero agradecimento por tudo.

Agradeço também aos amigos Gabriel Ribeiro e Roger Alves, parceiros inseparáveis desta longa jornada. Juntos, formamos o famoso Quarteto Fantástico, um laço marcado por união, força, risadas, quedas e superações que moldaram nossa história. Cada momento vivido ao lado de vocês carrega um brilho especial, e essa amizade leal e verdadeira levarei comigo por toda a vida.

Ao meu amigo Geraldo Rodrigues, minha gratidão sincera por fazer parte dessa caminhada. Mesmo à distância, sua presença, apoio e palavra amiga sempre encontraram um jeito de chegar até mim. Nada, nem o tempo nem os quilômetros, diminuíram a importância do que compartilhamos. Agradeço também ao meu grande amigo Antonio Heitor, pelo amparo constante, pelas risadas que iluminaram os dias difíceis e pela presença leal que acompanhou e fortaleceu cada etapa desta jornada. Ao meu amigo Atos Brandão, agradeço pelo companheirismo de longa data, pela parceria constante e por todas as conversas, risadas e incentivos que fortaleceram minha caminhada

Aos meus professores, registro minha profunda gratidão por todo conhecimento compartilhado, pelos ensinamentos que moldaram não apenas meu percurso acadêmico, mas também minha formação pessoal e profissional. De forma especial, e como verdadeira fonte de inspiração, agradeço ao professor Ramasio Melo, a quem carinhosamente chamo de meu mestre. Obrigado por enxergar em mim um potencial que, muitas vezes, eu mesmo não via; por suas palavras firmes e inspiradoras, que reacenderam minha fé nos momentos de dúvida; pela paciência constante e pela generosidade em compartilhar saberes. Cada projeto desenvolvido ao seu lado foi uma experiência audaciosa, capaz de ampliar meu olhar sobre a ciência, sobre o mundo e, sobretudo, sobre quem eu sou e posso me tornar.

Agradeço ainda aos professores Ennio Willian e Rogério Pereira, pelos momentos inspiradores, pela confiança em meu trabalho e pelas mentorias e oportunidades de desenvolvermos grandes projetos em conjunto. À professora Jeane Cristina, minha gratidão mais calorosa e virtuosa. Desde o primeiro instante, sua presença irradiou acolhimento, ternura e cuidado. Em cada orientação, em cada palavra de incentivo e em cada gesto de atenção, senti-me não apenas como aluno, mas como alguém recebido com o afeto e a proteção de um filho. Seu carinho, seus ensinamentos e sua sensibilidade marcaram minha trajetória de um modo tão especial que levarei sua presença comigo por toda a vida.

À professora Kênia Maria, agradeço pelo jeito leve e divertido de ensinar, pelas palavras sábias, pelos conselhos sobre crescimento pessoal e pesquisa, e por todo auxílio nos momentos em que mais precisei.

Enfim, deixo registrado meu agradecimento a todos que fizeram parte da minha caminhada. Cada gesto, palavra, incentivo e presença contribuiu para que este sonho se tornasse possível. Levo comigo cada pessoa, cada memória e cada aprendizado. A todos, expresso minha mais sincera gratidão por tornarem tudo tão especial.

“Se você vencer, você vive. Se você perder, você morre. Se você não lutar, você não pode vencer. Então lute, lute, lute.”

Eren Yeager

RESUMO

Este trabalho investiga o desafio de recuperar informações acadêmicas com precisão a partir de documentos institucionais extensos, cenário no qual assistentes virtuais costumam apresentar dificuldades, especialmente diante de estruturas tabulares. O objetivo consistiu em desenvolver, implementar e avaliar um agente conversacional aplicado à gestão acadêmica do Instituto Federal do Tocantins – *Campus Araguatins* (IFTO), capaz de consultar normas, currículos e regulamentos com maior confiabilidade. A pesquisa adotou a metodologia *Design Science Research* e propôs uma arquitetura baseada em segmentação por elementos documentais, busca híbrida e *reranking*. A validação empírica ocorreu por meio de interações reais, totalizando 450 mensagens enviadas por usuários humanos, 228 respostas automáticas e 224 feedbacks válidos. Os resultados demonstraram elevada eficácia do artefato, alcançando 83,9% de precisão nas consultas avaliadas e um Índice de Satisfação do Usuário de 4,32 em uma escala de 1 a 5. Além disso, o sistema apresentou desempenho consistente em consultas que envolvem dados tabulares, tradicionalmente problemáticas em sistemas automatizados. Conclui-se que o agente desenvolvido oferece uma contribuição metodológica relevante para a aplicação de Inteligência Artificial na gestão acadêmica, consolidando-se como uma solução replicável para instituições que lidam com grande volume de documentos complexos.

Palavras-chave: Inteligência Artificial. Gestão Acadêmica. Agente Conversacional. RAG Híbrido com Reranking. Modelos de Linguagem (LLMs).

ABSTRACT

This study investigates the challenge of accurately retrieving academic information from extensive institutional documents, a context in which virtual assistants often exhibit limitations, particularly when handling tabular structures. The objective was to develop, implement, and evaluate a conversational agent applied to academic management at the Federal Institute of Tocantins – *Campus Araguatins* (IFTO), capable of consulting regulations, curricula, and institutional guidelines with increased reliability. The research adopted the *Design Science Research* methodology and proposed an architecture based on document-element segmentation, hybrid search, and *reranking*. Empirical validation was conducted through real user interactions, totaling 450 messages sent by human users, 228 automated responses, and 224 valid feedback instances. The results demonstrated high effectiveness of the artifact, achieving 83.9% accuracy in the evaluated queries and a User Satisfaction Index of 4.32 on a 1–5 scale. In addition, the system exhibited consistent performance in queries involving tabular data, traditionally problematic for automated systems. It is concluded that the developed agent provides a relevant methodological contribution to the application of Artificial Intelligence in academic management, establishing itself as a replicable solution for institutions dealing with large volumes of complex documentation.

Keywords: Artificial Intelligence. Academic Management. Conversational Agent. Hybrid RAG with Reranking. Language Models (LLMs).

LISTA DE ILUSTRAÇÕES

Quadro 1 - Composição do Corpus Documental Institucional do IFTO.	18
Quadro 2 - Estratégias de Prompting Utilizadas no fluxo.	23
Figura 1 - Fluxo de Ingestão e Vetorização de Documentos.	25
Figura 2 - Fluxo de Inferência RAG com Busca Híbrida e Reranking.	25
Figura 3 - Comparativo de Desempenho entre os dois Assistentes Virtuais.	26
Figura 4 - Gráficos de Barras Comparativos - Tempo, ISU e Taxa de Acerto.	27
Figura 5 - Gráficos de Pizza Comparativos - Distribuição Temática.	28
Figura 6 - Interação do assistente em consulta normativa sobre ética em pesquisa.	29

LISTA DE SIGLAS

AI – Artificial Intelligence

API – Application Programming Interface

CEP – Comitê de Ética em Pesquisa

CEUA – Comissão de Ética no Uso de Animais

DSR – Design Science Research

GPT – Generative Pre-trained Transformer

IFTO – Instituto Federal do Tocantins

ISU – Índice de Satisfação do Usuário

JSON – JavaScript Object Notation

LGPD – Lei Geral de Proteção de Dados

LLM – Large Language Model

ODP – Organização Didático-Pedagógica

PDI – Plano de Desenvolvimento Institucional

PIBID – Programa Institucional de Bolsas de Iniciação à Docência

PPC – Projeto Pedagógico de Curso

RAG – Retrieval-Augmented Generation

TA – Taxa de Acerto

TI – Tecnologia da Informação

UUID – Universally Unique Identifier

VBA – Visual Basic for Applications

SUMÁRIO

1 INTRODUÇÃO 15

2 TRABALHOS RELACIONADOS 16

2.1 LLMs e Chatbots na Educação 16

2.2 Desafios do RAG em Documentos Estruturados 16

2.3 Reranking como Mecanismo de Precisão 17

3 MATERIAIS E MÉTODOS 17

3.1 Protocolo De Pré-Processamento E Higienização Do Corpus 17

3.2 Algoritmo de Reclassificação Híbrida (Hybrid Reranking) 21

3.3 Engenharia de Prompts e Instruções Sistêmicas 22

4 RESULTADOS E DISCUSSÕES 24

4.1 Eficiência Computacional e Análise de Custos 30

5 CONSIDERAÇÕES FINAIS 30

REFERÊNCIAS 32

1 INTRODUÇÃO

A transformação digital no ensino superior transcendeu a simples informatização de processos para constituir ecossistemas complexos, nos quais a Inteligência Artificial (IA) atua como vetor central de inovação (Zhalnina et al., 2024). Nesse cenário, a aplicação de Grandes Modelos de Linguagem (*Large Language Models – LLMs*) integrados a arquiteturas de Geração Aumentada por Recuperação (*Retrieval-Augmented Generation – RAG*) consolidou-se como o paradigma dominante (Genesis, 2025).

Sendo assim, essa abordagem é fundamental para mitigar alucinações em respostas geradas por IA, garantindo maior confiabilidade no suporte educacional, conforme apontado por Swacha e Gracel (2025) em sua revisão sistemática sobre chatbots na educação. Contudo, a literatura recente indica que implementações tradicionais, denominadas *Naive RAG*, enfrentam severas limitações ao processar documentos institucionais densos. Essa dificuldade é crítica em arquivos ricos em estruturas tabulares e hierarquias semânticas, formatos predominantes no ambiente burocrático universitário.

Dessa forma, o problema central desta pesquisa reside na ineficácia dos algoritmos de segmentação (*chunking*) baseados em contagem fixa de caracteres e na recuperação puramente vetorial ao lidar com dados estruturados. Tais métodos frequentemente dissociam cabeçalhos de seus respectivos valores, comprometendo a integridade da informação recuperada. Um estudo preliminar conduzido pelos próprios autores deste trabalho no contexto do Instituto Federal do Tocantins (IFTO), que estabelece a linha de base (*baseline*) desta investigação, evidenciou empiricamente essas falhas.

Na ocasião, utilizando uma arquitetura RAG padrão, foram obtidos apenas 74,1% de acerto, com erros concentrados em consultas que demandavam a recuperação precisa de dados curriculares tabulados. Esse lapso de desempenho corrobora as descobertas de Neumann et al. (2025), que destacam a necessidade imperativa de mecanismos mais robustos de verificação de fatos em chatbots acadêmicos.

Diante desse cenário, formula-se a seguinte Questão de Pesquisa: Como uma arquitetura de RAG Avançado, fundamentada em um reconhecimento de estrutura e busca híbrida com reclassificação (*reranking*), pode superar as limitações de precisão na recuperação de informações em documentos normativos do IFTO? A hipótese central é que a preservação da integridade semântica de tabelas durante a ingestão de dados, combinada a estratégias de busca que mesclam vetores semânticos e palavras-chave exatas, reduz drasticamente a taxa de alucinação do modelo em consultas administrativas complexas.

Para responder a essa questão, o presente trabalho propõe o desenvolvimento e a

avaliação de um agente conversacional focado nos documentos do Curso de Licenciatura em Computação, incluindo Projetos Pedagógicos de Curso (PPC), regulamentos de estágio e normas internas. A abordagem técnica baseia-se no modelo Re²G (*Retrieve, Rerank, Generate*) discutido por Glass et al. (2022), escolhido por sua capacidade de filtrar ruídos irrelevantes através de um estágio de reclassificação, priorizando a recuperação de dados exatos.

Além da precisão técnica, a pesquisa aborda desafios éticos críticos identificados por Yan et al. (2023), como a transparência algorítmica e a beneficência. A solução é desenhada sob o amparo da Lei Geral de Proteção de Dados (LGPD), respondendo à necessidade de adequação dos Institutos Federais apontada por Rojas e Medeiros (2021), que identificaram a gestão de dados pessoais e a mudança cultural institucional como pontos críticos nessas instituições.

Portanto, o objetivo geral deste trabalho consiste em desenvolver, implementar e avaliar um agente conversacional baseado em RAG, capaz de elevar a precisão da recuperação de informações acadêmicas para níveis superiores ao estado da arte local. Especificamente, busca-se validar se a combinação de *chunking* semântico estruturado e heurísticas de reclassificação com reforço de termos (técnica denominada neste estudo como *Boost*) é capaz de sanar as deficiências de recuperação em dados tabulares identificadas no *baseline*. Desta forma, o estudo oferece uma contribuição metodológica replicável para a aplicação de Inteligência Artificial em domínios administrativos de alta complexidade documental.

2 TRABALHOS RELACIONADOS

A integração de Grandes Modelos de Linguagem (LLMs) em ambientes educacionais tem sido objeto de intensa investigação, com foco predominante na personalização da aprendizagem e na automação de tutoria.

2.1 LLMs e Chatbots na Educação Zhalnina et al. (2024) discutem o papel transformador da IA no ensino superior, destacando que a digitalização transcendeu a infraestrutura básica para alcançar a camada cognitiva da gestão educacional. No entanto, Lee e Moore (2024), em revisão sistemática, alertam que a maioria das aplicações atuais foca no feedback pedagógico (correção de tarefas), deixando uma lacuna significativa na assistência administrativa e burocrática, área que consome tempo substancial de coordenadores e docentes.

2.2 Desafios do RAG em Documentos Estruturados A arquitetura RAG (Retrieval-Augmented Generation) surgiu como solução para a desatualização dos LLMs. Contudo, implementações padrão (Naive RAG) falham ao processar documentos complexos. Jimeno-

Yepes et al. (2024) demonstraram que a segmentação (chunking) baseada apenas em contagem de caracteres destrói a semântica de tabelas financeiras, um cenário análogo às matrizes curriculares presentes nos Projetos Pedagógicos de Curso (PPC). O estudo de Song (2025) reforça que a preservação hierárquica de cabeçalhos é o fator determinante para a precisão em recuperações tabulares.

2.3 Reranking como Mecanismo de Precisão Para mitigar a recuperação de contextos irrelevantes, Glass et al. (2022) propuseram o paradigma Re²G (Retrieve, Rerank, Generate), validando o uso de reclassificadores para filtrar ruído. Recentemente, Wei et al. (2024) e Mala et al. (2025) confirmaram que arquiteturas híbridas (vetorial + lexical) combinadas com reranking reduzem drasticamente as alucinações, pois impedem que o modelo gerador receba informações contraditórias na janela de contexto. Este trabalho aplica esses conceitos teóricos especificamente no domínio da gestão acadêmica brasileira, aderente às normas do IFTO.

3 MATERIAIS E MÉTODOS

Inicialmente, para atender aos objetivos propostos e responder à questão de pesquisa, este estudo adota a abordagem metodológica da Design Science Research (DSR). Conforme Hevner et al. (2004), esta metodologia não se limita à criação de um artefato tecnológico, mas investiga a eficácia da solução na resolução de problemas práticos em contextos complexos, iterando através de ciclos de design, desenvolvimento e avaliação. A arquitetura metodológica foi desenhada para superar limitações previamente identificadas no contexto institucional. Um diagnóstico preliminar realizado no IFTO em 2024, no contexto de testes internos com uma arquitetura RAG padrão, estabeleceu o baseline desta investigação. Os resultados, acurácia de 74,1% em consultas envolvendo tabelas foram disponibilizados para este estudo e servem como referência comparativa.

3.1 Protocolo De Pré-Processamento E Higienização Do Corpus

Em primeira análise, o corpus documental da pesquisa é composto por 18 documentos institucionais oficiais do Instituto Federal do Tocantins (IFTO), abrangendo o Projeto Pedagógico do Curso (PPC), o Plano de Desenvolvimento Institucional (PDI) e ao Regulamento da Organização Didático-Pedagógica dos Cursos de Graduação Presenciais do IFTO (ODP), sendo que o documento mais extenso possui 276 páginas. A constituição da base de conhecimento (Knowledge Base) envolveu a curadoria de 18 documentos institucionais oficiais

do IFTO, selecionados para cobrir todas as dimensões da vida discente, desde a infraestrutura física até as normas de pesquisa e ética. A diversidade desses arquivos exigiu um mapeamento detalhado para garantir que o sistema tivesse cobertura normativa completa.

O Quadro 1 apresenta o inventário integral dos documentos processados pelo sistema, categorizados por sua natureza administrativa e pedagógica.

Quadro 1 – Composição do Corpus Documental Institucional do IFTO

Categoria	Documentos Oficiais	Descrição do Conteúdo
Estruturantes e Institucionais	1. PDI - Plano de Desenvolvimento Institucional 2. PPI - Projeto Político Pedagógico Institucional (2020-2024) 3. PDTI - Plano Diretor de TI (2023) 4. Regimento Geral do IFTO (2019)	Instrumentos de planejamento estratégico que definem a identidade institucional, a hierarquia organizacional, as políticas de gestão macro e os direitos fundamentais da comunidade acadêmica.
Pedagógicos e do Curso	5. PPC - Projeto Pedagógico do Curso de Licenciatura em Computação (2023) 6. Regulamento da Organização Didático-Pedagógica (ODP) 7. Diretrizes para Elaboração de TCC 8. Regulamento do Estágio Supervisionado	Documentos críticos contendo matrizes curriculares, ementas, sistemas de avaliação e fluxos acadêmicos específicos. São as fontes primárias para dúvidas sobre "o que estudar" e "como ser aprovado".
Pesquisa, Extensão e Inovação	9. Regulamento das Atividades de Pesquisa 10. Regulamento de Iniciação Científica 11. Regulamento das	Normativas que regem a produção científica, bolsas, editais de fomento e a integração do aluno com a comunidade externa, essenciais para a formação complementar.

	Atividades de Extensão 12. Curricularização da Extensão 13. Regimento Interno do PIBID	
Conduta, Ética e Infraestrutura	14. Regulamento Disciplinar Discente 15. Código de Conduta Ética do IFTO 16. Comitê de Ética no Uso de Animais (CEUA) 17. Normas do Laboratório de Informática (NI-LI) 18. Documentos de Infraestrutura	Regras de convivência, diretrizes éticas para pesquisa com seres vivos e normas operacionais para o uso seguro de laboratórios e espaços físicos do campus.

Fonte: Elaborado pelos autores, (2025).

A heterogeneidade desses arquivos, que alternam entre texto corrido denso (como no PDI e PPI) e tabelas complexas de matriz curricular (no PPC e ODP), justificou a adoção do algoritmo de limpeza via VBA. Conversores padrão falharam em manter a integridade das linhas e colunas, especialmente nas normas de laboratório e cronogramas de extensão, o que exigiu a padronização prévia descrita a seguir.

Este processo envolve a remoção de cabeçalhos repetitivos e logotipos que geram ruído semântico, bem como a reescrita e simplificação de tabelas complexas para um formato alinhado. Esta etapa de preparação manual é crucial, pois, conforme apontado por Yan et al. (2023), a qualidade dos dados de entrada impacta diretamente a performance e a confiabilidade das respostas geradas por modelos de linguagem em contextos educacionais.

A constituição da base de conhecimento (*Knowledge Base*) envolveu a curadoria de 18 documentos institucionais heterogêneos, que variam desde regulamentos normativos estritos (como o ODP e o Regimento Geral) até documentos programáticos extensos, como o Projeto Pedagógico de Curso (PPC) e o Plano de Desenvolvimento Institucional (PDI). A análise preliminar desses arquivos, que somam centenas de páginas, revelou que a extração direta de

texto via bibliotecas padrão (como *PyPDF* ou *pdf.js*) resultava em ruído significativo, misturando cabeçalhos repetitivos e rodapés com o conteúdo útil, além de desconfigurar completamente as tabelas complexas.

Para mitigar esse problema e garantir a qualidade da ingestão (*Garbage In, Garbage Out*), desenvolveu-se uma rotina de automação prévia utilizando *Visual Basic for Applications* (VBA) no ambiente Microsoft Word. O algoritmo, denominado *TabulacaoCleanDocument*, foi projetado para normalizar a estrutura dos arquivos antes da conversão para *Markdown*. A rotina executa três operações críticas sequenciais: (i) a varredura e remoção iterativa de todos os cabeçalhos e rodapés das seções, eliminando repetições que poderiam enviesar a busca vetorial; (ii) a conversão programática de objetos de tabela em texto estruturado, utilizando tabulações (*wdSeparateByTabs*) como delimitadores lógicos, o que preserva a relação linha-coluna legível para o LLM; e (iii) a exclusão de elementos gráficos não textuais (imagens e formas) que não contribuem para a recuperação semântica neste escopo.

Dessa forma, essa etapa de higienização automatizada foi determinante para o sucesso da segmentação subsequente. Ao transformar tabelas complexas em estruturas de texto linearizadas por tabulação, facilitou-se a detecção de padrões pelo algoritmo de *chunking* baseado em *Regex*, assegurando que os dados tabulares fossem tratados como blocos de informação coesos e não fragmentados.

Após a curadoria, implementou-se um pipeline automatizado de ingestão de dados, mediado pela plataforma *low-code n8n*. Diferentemente das abordagens tradicionais de segmentação (*chunking*) por contagem fixa de caracteres, que tendem a fragmentar tabelas e dissociar cabeçalhos de seus valores, este trabalho desenvolveu uma estratégia denominada *Structure-Aware Chunking* (Segmentação com Reconhecimento de Estrutura).

Utilizando um algoritmo em JavaScript personalizado baseado em Expressões Regulares (*Regex*), o sistema processa os documentos em formato *Markdown*, identificando e preservando blocos de tabelas (detectados pela sintaxe `|---`) como unidades semânticas indivisíveis. Esta abordagem impede que informações críticas, como a carga horária total do curso, sejam separadas de seus rótulos descritivos.

A escolha por este método fundamenta-se nas descobertas de Jimeno-Yepes et al. (2024), que demonstraram que o *chunking* baseado em elementos estruturais supera significativamente as estratégias de tamanho fixo na recuperação de informações em relatórios

financeiros e técnicos densos. Adicionalmente, Song (2025) reforça que a representação adequada de cabeçalhos e estruturas hierárquicas em dados tabulares é essencial para melhorar a performance de sistemas *RAG*.

Para cada segmento processado, o sistema gera um *UUID* determinístico e vetoriza o conteúdo utilizando o modelo *text-embedding-3-small* da OpenAI, armazenando o vetor, o texto original e os metadados via requisição HTTP no banco de dados vetorial Qdrant. O fluxo inclui ainda uma etapa de idempotência, onde vetores antigos são deletados antes da reinserção, garantindo a consistência da base de conhecimento.

Por conseguinte, o momento da inferência, ou seja, a interação em tempo real com o usuário via WhatsApp, opera através de uma arquitetura de *RAG* composta por múltiplos estágios. Inicialmente, a consulta do usuário é processada por um agente intermediário (*Gemini 2.5 Flash*) responsável pela Expansão de Consulta (*Query Expansion*). Este agente possui instruções sistêmicas para inferir termos técnicos implícitos; por exemplo, ao detectar uma pergunta sobre "carga horária", o sistema enriquece a busca adicionando termos-chave obrigatórios.

3.2 Algoritmo de Reclassificação Híbrida (Hybrid Reranking)

Nesse contexto, a recuperação da informação ocorre através de uma estratégia de Busca Híbrida (*Hybrid Search*), que executa simultaneamente uma busca vetorial (semântica) e uma busca lexical (palavras-chave). Esta abordagem mitiga as limitações dos vetores em capturar termos exatos, como códigos de disciplinas ou números específicos. Os resultados provenientes desses dois caminhos são fundidos e submetidos a um algoritmo de reclassificação (*Reranking*) desenvolvido especificamente para este estudo.

A função de pontuação, denominada heurísticamente de *Boost*, atribui pesos dinâmicos aos documentos recuperados: uma pontuação base derivada da similaridade vetorial é acrescida de penalidades positivas caso o segmento seja identificado como uma tabela ou contenha os termos-chave exatos definidos na etapa de expansão. Esta lógica de reclassificação alinha-se ao modelo *Re²G* (*Retrieve, Rerank, Generate*) proposto por (Glass et al., 2022), que valida o uso de *rerankers* para filtrar documentos irrelevantes e priorizar passagens que maximizam a precisão da geração subsequente. Finalmente, os documentos reordenados são enviados ao modelo gerador (*Gemini 2.5 Pro*), que sintetiza a resposta final ao usuário. A principal inovação da arquitetura proposta reside no estágio de reclassificação (*Reranking*), implementado como um nó de processamento lógico intermediário (*Middleware*) entre a recuperação e a geração.

Diferentemente de abordagens que confiam puramente na similaridade do cosseno, este estudo implementou uma função de pontuação heurística personalizada (*Custom Scoring Function*) para refinar a relevância dos documentos recuperados.

O algoritmo desenvolvido opera através da ponderação de duas fontes de sinal: o Score Vetorial (proveniente do banco Qdrant, refletindo a similaridade semântica) e o Score Lexical (calculado localmente com base na presença exata de palavras-chave extraídas na etapa de expansão). A função de fusão atribui pesos distintos a esses componentes, priorizando a semântica ($sw_v = 0.6$) sobre a correspondência exata ($sw_l = 0.4$), mas utilizando a segunda como critério de desempate crítico para termos técnicos específicos. Além da fusão linear, o algoritmo aplica "Boosts" (incrementos de pontuação) baseados em metadados estruturais, conforme a lógica descrita a seguir:

1. **Deteção de Intenção Numérica:** O sistema analisa os *tokens* da consulta do usuário. Se forem detectados termos quantitativos (ex: "total", "carga horária", "duração") em conjunto com numerais, ativa-se um modo de precisão.
2. **Bônus Estrutural para Tabelas:** Segmentos de texto que foram identificados durante a ingestão como tabelas (marcados pela sintaxe Markdown |---|) recebem um multiplicador de relevância ($1.5 \times$) quando a intenção numérica está ativa. Isso força o sistema a priorizar a "fonte primária" do dado (a tabela) em detrimento de menções textuais esparsas.
3. **Ancoragem de Termos Críticos:** Foi implementada uma regra de domínio específica para garantir a recuperação da carga horária de integralização do curso. Documentos que contêm valores exatos críticos definidos na matriz curricular recebem um acréscimo aditivo na pontuação final, assegurando que a informação oficial do PPC prevaleça sobre aproximações.

Concluindo, essa abordagem híbrida permite que o sistema exiba um comportamento resiliente: ele compreende perguntas formuladas em linguagem natural (via vetores), mas não alucina números ou códigos, mantendo-se ancorado nos dados exatos (via palavras-chave e estrutura).

3.3 Engenharia de Prompts e Instruções Sistêmicas

A eficácia do agente conversacional não reside apenas na recuperação de dados, mas na capacidade de interpretar a intenção do usuário e formatar a resposta dentro dos limites institucionais. Para isso, em vez de instruções genéricas, foram desenvolvidas diretrizes

comportamentais estritas (*Behavioral Guidelines*) para cada estágio do *pipeline* no n8n, aplicando técnicas de *Prompt Engineering* focadas em delimitação de escopo e ancoragem semântica.

O Quadro 2 apresenta as estratégias lógicas implementadas nos agentes. Destaca-se o mecanismo de "Ancoragem Numérica" no Agente Expansor, desenhado para resolver a ambiguidade em consultas quantitativas, forçando o sistema a priorizar a recuperação de tabelas integrais em vez de fragmentos de texto isolados.

Quadro 2 – Estratégias de Prompting Utilizadas no fluxo

Agente	Estratégias e Restrições Implementadas
Expansor de Consulta (<i>Query Expander</i>)	1. Tradução de Domínio: Conversão de linguagem natural informal para terminologia técnica institucional (ex: tradução automática de siglas como "PPC" e "ODP"). 2. Ancoragem Numérica: Ativação de gatilhos lógicos quando termos quantitativos são detectados (ex: "carga horária"), forçando a inclusão de metadados de tabelas específicas na busca vetorial. 3. Expansão de Sinônimos: Enriquecimento da <i>query</i> com termos correlatos para aumentar o <i>recall</i> da busca híbrida.
Agente Gerador (<i>Response Generator</i>)	1. Restrição de Contexto (<i>Grounding</i>): Bloqueio total de conhecimento externo ao <i>corpus</i>, obrigando o modelo a responder exclusivamente com os fragmentos recuperados. 2. Preservação Estrutural: Diretriz mandatória para manter a formatação original de dados tabulares (Markdown) na resposta

	final. 3. Protocolo de Falha: Instrução explícita para admitir o desconhecimento ("Não encontrei a informação") em vez de inferir ou alucinar dados não presentes nos documentos.
--	--

Fonte: Elaborado pelos autores, (2025).

Concluindo, a validação do artefato seguiu um protocolo de Avaliação Centrada no Usuário (*Human-Centric Evaluation*). Conforme destacado por Abeysinghe e Circi (2024), a avaliação humana continua sendo um elemento indispensável na avaliação de chatbots, permitindo uma análise mais enriquecedora das respostas e da eficácia do sistema educativo. Diferente de métricas puramente automáticas, a avaliação humana é capaz de capturar nuances de utilidade pedagógica e clareza, essenciais em um chatbot educacional.

O desempenho foi mensurado através da coleta de feedbacks reais dos usuários (discentes e docentes) em uma escala Likert de 1 a 5, permitindo o cálculo da Taxa de Acerto (TA) e do Índice de Satisfação do Usuário (ISU). Em conformidade com os princípios éticos levantados por Yan et al. (2023), o sistema foi projetado para garantir a transparência algorítmica e a privacidade dos dados.

O tratamento das informações esteve em conformidade com a Lei Geral de Proteção de Dados (LGPD), uma necessidade crítica para instituições federais conforme apontado por Rojas e Medeiros (2021) que destacam a importância da adequação de processos e sistemas institucionais para a proteção de dados pessoais.

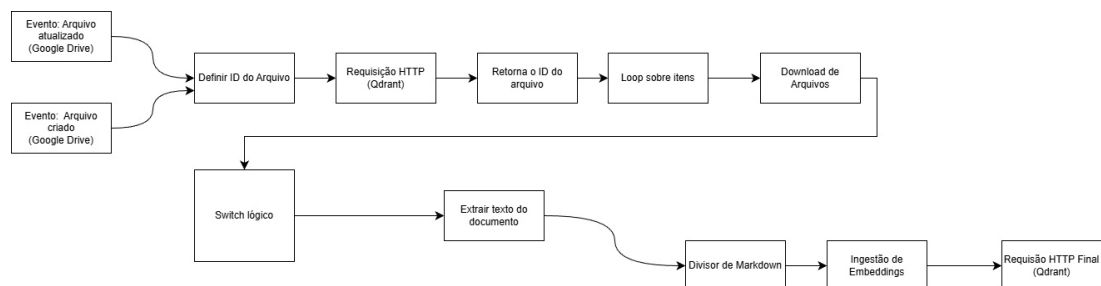
4 RESULTADOS E DISCUSSÕES

Em primeiro lugar, a apresentação dos resultados inicia-se pela demonstração do artefato tecnológico desenvolvido, cuja arquitetura foi desenhada para solucionar o problema da fragmentação de dados tabulares. A validação operacional dos fluxos de automação constitui o primeiro achado empírico deste estudo, confirmando a viabilidade técnica da abordagem proposta. Para operacionalizar a metodologia de Segmentação com reconhecimento de estrutura, foi desenvolvido um fluxo de ingestão de dados automatizado, conforme ilustrado na Figura 1. O diagrama evidencia a complexidade do tratamento dado aos documentos institucionais.

Diferente de pipelines lineares padrão, este fluxo incorpora nós de decisão e processamento de código personalizado.

O destaque técnico da Figura 1 reside no nó central "Divisor de Markdown". É neste estágio que o algoritmo baseado em *Regex* intercepta o texto bruto extraído do PDF e, antes da vetorização, identifica e preserva os blocos de tabelas (sintaxe `---`). O fluxo demonstra ainda a preocupação com a consistência da base de dados através dos nós de "Requisição HTTP" configurados para deletar vetores antigos antes da reinserção, garantindo que o assistente não recupere informações obsoletas de versões anteriores dos documentos.

Figura 1 - Fluxo de Ingestão e Vetorização de Documentos.

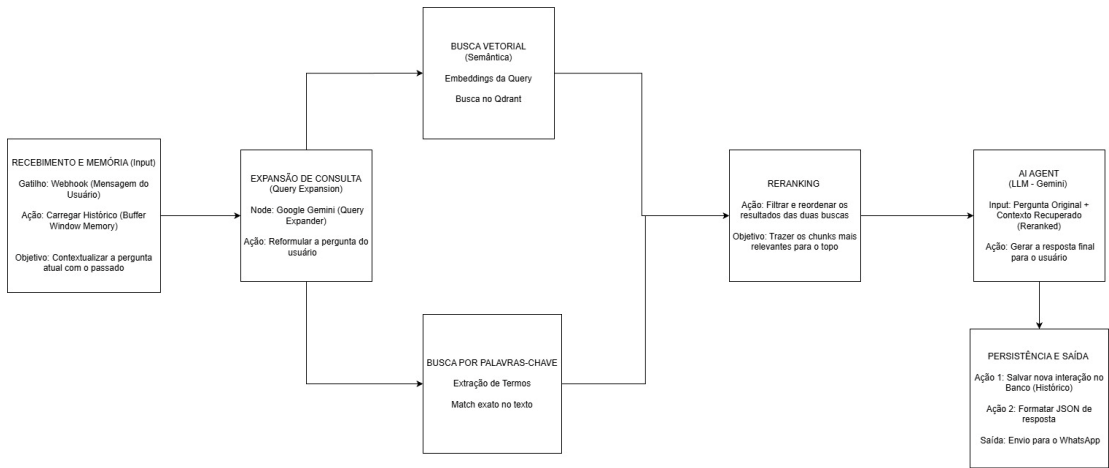


Fonte: Elaborado pelos autores, (2025).

Uma vez processados os dados, a interação com o usuário é gerida pelo fluxo de inferência apresentado na Figura 2. Este modelo conceitual materializa a arquitetura Re²G adaptada para o contexto do IFTO. A imagem detalha a execução paralela das estratégias de recuperação: o caminho superior executa a busca vetorial, enquanto o inferior executa a busca por palavras-chave.

Logo, o ponto crítico de inovação é visualizado no nó "*Reranking*", posicionado após a fusão dos resultados. É neste componente que a heurística de impulso é aplicada, reordenando os fragmentos recuperados com base na presença de estruturas tabulares e termos exatos. O fluxo encerra-se com o nó "*AI Agent*", que recebe o contexto já refinado para gerar a resposta final, evidenciando que a "inteligência" do sistema não reside apenas no modelo de linguagem, mas na engenharia de recuperação que o antecede.

Figura 2 - Fluxo de Inferência RAG com Busca Híbrida e Reranking.



Fonte: Elaborado pelos autores, (2025).

Após a validação da arquitetura, a análise dos dados obtidos a partir da interação dos usuários com o artefato desenvolvido revela um desempenho superior ao baseline estabelecido, confirmando a eficácia da arquitetura de RAG proposta para o contexto administrativo do IFTO. A avaliação quantitativa da eficácia do sistema, mensurada pela Taxa de Acerto (TA), demonstrou um incremento substancial de 9,8 pontos percentuais em relação ao modelo anterior, elevando a eficácia de 74,1% para 83,9%. Esse desempenho foi aferido a partir de um conjunto total de 678 mensagens trocadas no sistema, envolvendo a participação de quatro discentes e dois docentes. Entre essas interações, 224 receberam avaliações válidas na escala Likert (1 a 5), formando o corpus analítico utilizado para calcular as métricas de precisão e satisfação do usuário.

Figura 3 - Comparativo de Desempenho entre os dois Assistentes Virtuais.

DIMENSÃO	MÉTRICA	RESULTADO
EFICÁCIA	Taxa de Acerto (TA)	74,1%
SATISFAÇÃO	Índice de Satisfação (ISU)	3,87 (escala 1-5)

DIMENSÃO	MÉTRICA	RESULTADO
EFICÁCIA	Taxa de Acerto (TA)	83,9%
SATISFAÇÃO	Índice de Satisfação (ISU)	4,32 (escala 1-5)

Fonte: Elaborado pelos autores, (2025).

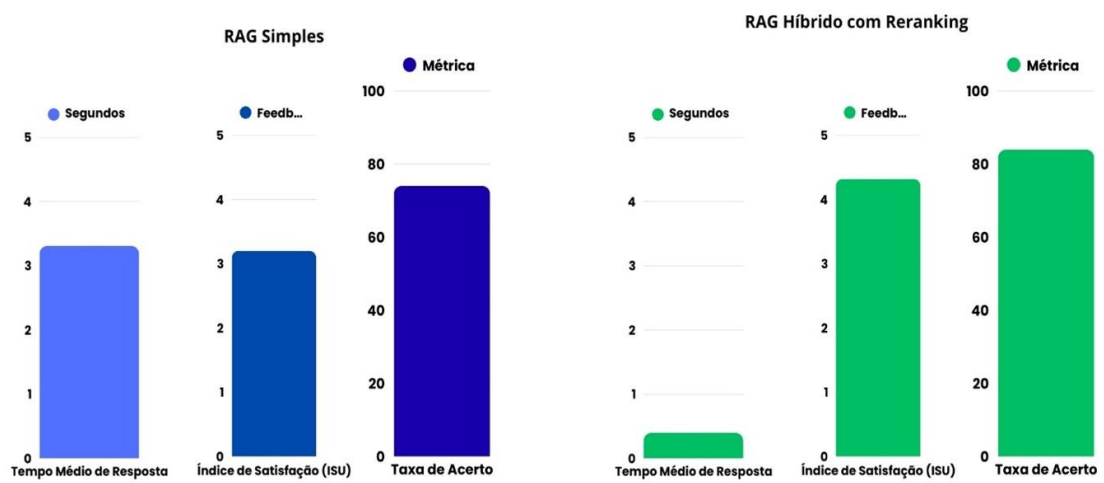
Consequentemente, a superioridade da abordagem híbrida com reclassificação (*reranking*) observada neste estudo corrobora as descobertas recentes da literatura especializada. Estudos de Wang et al. (2024) e (Mala et al., 2025) demonstram que recuperadores híbridos reduzem significativamente as taxas de alucinação ao combinar a precisão da busca lexical (palavras-chave) com a compreensão semântica dos vetores densos. No contexto do IFTO, essa estratégia permitiu que termos exatos críticos, como carga horária ou códigos de disciplinas, fossem ancorados corretamente. A aplicação de algoritmos de *reranking*, validados por (Wei et al., 2024) como eficazes na redução de informações contraditórias, garantiu que os documentos oficiais fossem priorizados sobre menções incidentais.

Outro aspecto técnico relevante refere-se à eficiência computacional e à economia de *tokens*. A arquitetura proposta implementa uma estratégia de uso racional da janela de contexto do LLM. Ao utilizar o modelo *Gemini 2.5 Flash* na etapa de Expansão de Consulta (agente intermediário), o sistema aproveita o baixo custo por token dessa versão para realizar tarefas de reformulação semântica, reservando o modelo mais robusto (*Gemini 2.5 Pro*) apenas para a geração da resposta final. Além disso, o estágio de *Reranking* atua como um filtro econômico crítico. No fluxo desenvolvido no n8n, embora a busca inicial no Qdrant recupere até 50 fragmentos candidatos, o algoritmo de reordenação seleciona e envia para o *AI Agent* apenas os 12 fragmentos (*chunks*) de maior relevância (*top_k=12*).

Assim, essa filtragem reduz drasticamente a quantidade de tokens de entrada (*input tokens*) processados pelo modelo gerador, diminuindo a latência da inferência e o custo operacional da API, sem sacrificar a qualidade da informação, uma vez que o conteúdo irrelevante é descartado antes de chegar ao LLM. Essa otimização é vital para a sustentabilidade do projeto, considerando que documentos como o PPC possuem estruturas extensas que, se enviadas integralmente ao *prompt*, consumiriam rapidamente a cota de tokens e aumentariam o tempo de resposta percebido pelo usuário.

Paralelamente à eficácia técnica, a análise da experiência do usuário revelou um Índice de Satisfação (ISU) de 4,32 em uma escala de 5 pontos. Este valor situa-se acima da média reportada na literatura para chatbots educacionais, que oscila entre 3,5 e 4,2 (Sáiz-Manzanares et al., 2023; Kayali et al., 2023). A distribuição dos feedbacks (Figura 4) evidencia uma melhoria consistente não apenas na precisão, mas na percepção de valor pelo usuário.

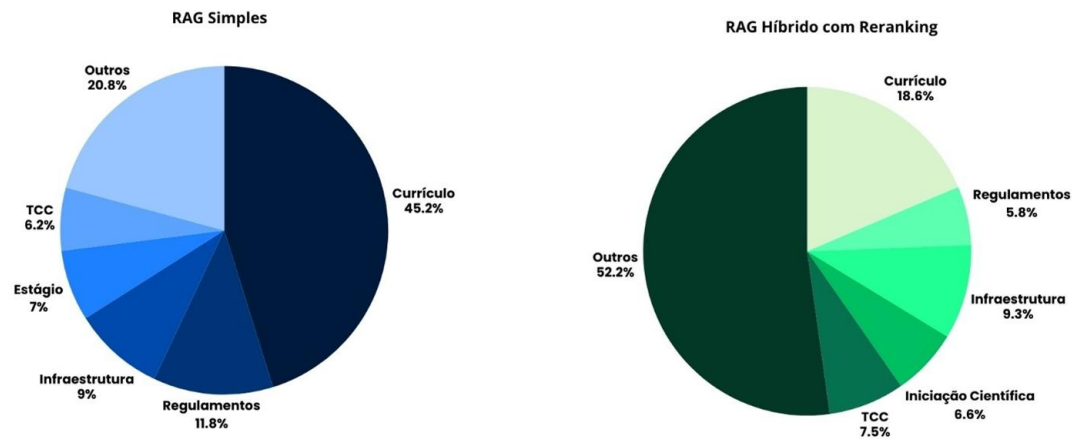
Figura 4 - Gráficos de Barras Comparativos - Tempo, ISU e Taxa de Acerto.



Fonte: Elaborado pelos autores, (2025).

Sendo assim, a análise dos padrões de demanda, ilustrada na Figura 5, revela uma mudança interessante no perfil de uso. Enquanto no baseline as perguntas concentravam-se massivamente em "Currículo" (45,2%), na nova arquitetura houve uma diversificação notável, com a categoria "Outros" representando 52,2% das interações. Isso sugere que a maior capacidade de compreensão do assistente encorajou os discentes a explorarem uma gama mais ampla de tópicos, saindo do óbvio. Ainda assim, temas críticos que dependem de dados tabulares, como "Currículo" (18,6%), "TCC" (7,5%) e "Infraestrutura" (9,3%), mantiveram relevância, confirmando a importância da Segmentação com reconhecimento de estrutura para atender a essas demandas específicas.

Figura 5 - Gráficos de Pizza Comparativos - Distribuição Temática.

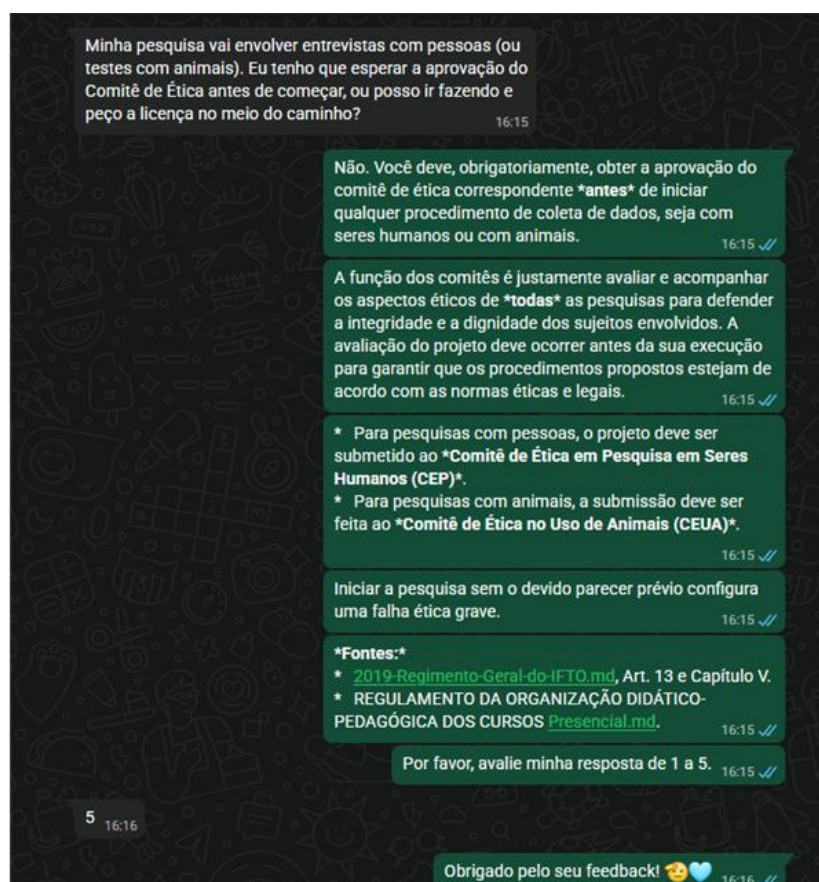


Fonte: Elaborado pelos autores, (2025).

Ademais, com o objetivo de demonstrar a capacidade do sistema em responder consultas normativas complexas em um contexto real de uso, a Figura 6 apresenta um excerto de uma conversa realizada via WhatsApp. Nessa interação, observa-se o assistente respondendo a uma dúvida relacionada aos procedimentos éticos necessários para a execução de pesquisas envolvendo seres humanos ou animais, tema que exige a correta interpretação e correlação de diferentes regulamentos institucionais. O sistema fornece uma resposta precisa e coerente, destacando a obrigatoriedade da aprovação prévia pelos comitês de ética competentes (CEP ou CEUA), conforme as normas vigentes do IFTO.

Em conclusão, a fluidez da linguagem e a adequação normativa da resposta evidenciam a eficácia da união entre mecanismos de recuperação de informações documentais e a capacidade generativa do modelo de linguagem, resultando em instruções claras, corretas e alinhadas ao contexto acadêmico institucional. A avaliação positiva do usuário (nota máxima) reforça a aceitabilidade e a confiabilidade percebida do sistema em situações reais de consulta acadêmica.

Figura 6 - Interação do assistente em consulta normativa sobre ética em pesquisa.



Fonte: Elaborado pelos autores, (2025).

Por último, é importante ressaltar que a compreensão do papel do *Reranking* na redução de alucinações requer evidências empíricas mais consistentes. A baixa taxa de feedbacks negativos sugere que os mecanismos de segurança funcionaram adequadamente, evitando a frustração gerada por alucinações confiantes e alinhando-se às recomendações éticas de Yan et al. (2023) sobre a beneficência em sistemas de IA.

4.1 Eficiência Computacional e Análise de Custos

A viabilidade de implantação do sistema em larga escala depende diretamente da eficiência no consumo de recursos computacionais. A análise dos *logs* de execução na plataforma n8n demonstrou que a arquitetura de *Reranking* atua como um filtro econômico essencial.

Enquanto a busca inicial no banco vetorial (Qdrant) recupera até 50 fragmentos candidatos para garantir cobertura (*recall*), o algoritmo de reclassificação seleciona apenas os 12 fragmentos mais relevantes (*top_k=12*) para compor a janela de contexto do Agente Gerador. Essa redução seletiva diminui em aproximadamente 76% o volume de *tokens* de entrada (*input tokens*) processados pelo modelo mais oneroso (Gemini 2.5 Pro).

Considerando o modelo de precificação atual das APIs, essa estratégia de "funil" não apenas reduz a latência da inferência final, mas viabiliza a sustentabilidade financeira do projeto. O uso do modelo *Gemini 2.5 Flash* (mais rápido e econômico) para a tarefa intermediária de expansão de consulta, reservando o modelo *Pro* apenas para a síntese final, demonstra um equilíbrio eficaz entre custo e inteligência, alinhando-se às práticas de *Green AI* ao reduzir o processamento desnecessário de dados.

5 CONSIDERAÇÕES FINAIS

O presente trabalho atingiu plenamente seu objetivo geral ao desenvolver, implementar e validar um agente conversacional baseado em arquitetura *RAG* capaz de elevar a precisão da recuperação de informações acadêmicas no contexto do Instituto Federal do Tocantins – *campus* Araguatins. Os resultados empíricos demonstraram que a articulação entre técnicas de Segmentação com reconhecimento de estrutura e estratégias de *reranking* híbrido permitiu superar de forma significativa as limitações observadas no modelo de linha de base, elevando a Taxa de Acerto de 74,1% para 83,9% e o Índice de Satisfação do Usuário de 3,87 para 4,32.

Assim, confirma-se a hipótese central da pesquisa: a preservação da integridade semântica de dados tabulares durante a ingestão constitui um dos principais fatores responsáveis pela mitigação de alucinações em documentos institucionais complexos. Além disso, a principal contribuição científica deste estudo reside na sistematização de uma

metodologia replicável para o tratamento de documentos institucionais não estruturados.

Enquanto a literatura predominante concentra-se na otimização de modelos de linguagem (*LLMs*), esta pesquisa evidencia que a engenharia de recuperação (*Retrieval Engineering*), especialmente a segmentação baseada em estrutura e o uso de busca híbrida pode ser ainda mais determinante para assegurar a fidelidade factual em domínios regulatórios. Além do avanço técnico, os resultados indicam que a ferramenta possui potencial para reduzir a sobrecarga administrativa enfrentada por docentes e coordenadores.

Conforme apontado por Lee et al. (2024) e Khairullah et al. (2025), a automação de tarefas rotineiras, como o esclarecimento de dúvidas sobre currículos e normativas, contribui para otimizar processos internos e melhorar o fluxo de trabalho acadêmico, favorecendo um ambiente institucional mais organizado e eficiente. Apesar dos resultados promissores, o estudo apresenta limitações importantes. A dependência de modelos proprietários disponibilizados via API (OpenAI/Google) implica custos operacionais contínuos e suscita questionamentos acerca da soberania dos dados no longo prazo.

Investigações recentes de Ateia et al. (2024) e Zolfaghari et al. (2024) demonstram que modelos de código aberto, como *Llama 3* e *Mistral*, já alcançam desempenho competitivo ao *GPT-4* em tarefas sensíveis à privacidade quando adequadamente ajustados. Migrar para uma infraestrutura local poderia eliminar riscos de exposição indevida de dados, fortalecendo a aderência aos princípios da LGPD.

Outra limitação identificada refere-se à incapacidade do sistema atual de interpretar elementos visuais, como fluxogramas, organogramas e gráficos estatísticos, presentes em parte relevante dos documentos institucionais. Diante disso, sugerem-se, como diretrizes para pesquisas futuras, a exploração de arquiteturas de *RAG* Multimodal. Conforme discutem Yang et al. (2025) e Wang et al. (2024), a incorporação de capacidades de processamento visual em sistemas *RAG* representa a próxima fronteira para a compreensão aprofundada de documentos educacionais ricos em elementos gráficos.

Recomenda-se, adicionalmente, a realização de estudos longitudinais que avaliem o impacto da adoção desse tipo de ferramenta na redução do burnout da equipe administrativa, considerando que, segundo a literatura, os efeitos positivos da IA na saúde ocupacional dependem não apenas da tecnologia em si, mas da forma como sua implementação é conduzida no ambiente institucional.

REFERÊNCIAS

ABEYSINGHE, Bhashithe; CIRCI, Ruhan. **The Challenges of Evaluating LLM Applications:** An Analysis of Automated, Human, and LLM-Based Approaches. arXiv:2406.03339v2 [cs.CL], 2024.

ATEIA, Samy; KRUSCHWITZ, Udo. **Can Open-Source LLMs Compete with Commercial Models?:** Exploring the Few-Shot Performance of Current GPT Models in Biomedical Tasks. 2024. 21 p. Preprint (arXiv:2407.13511). Disponível em: <https://arxiv.org/abs/2407.13511>. Acesso em: 26 nov. 2025.

GENESIS, J. **Retrieval-Augmented Text Generation:** Methods, Challenges, and Applications. [s. l.], 2025.

GLASS, Michael *et al.* **Re²G:** Retrieve, Rerank, Generate. 2022. 15 p. Preprint (arXiv:2207.06300). Disponível em: <https://arxiv.org/abs/2207.06300>. Acesso em: 25 nov. 2025.

HEVNER, A. R.; MARCH, S. T.; PARK, J.; RAM, S. **Design Science in Information Systems Research.** MIS Quarterly, v. 28, n. 1, p. 75-105, 2004. Disponível em: <https://doi.org/10.2307/25148625>. Acesso em: 6 dez. 2025.

JIMENO YEPES, Antonio *et al.* **Financial Report Chunking for Effective Retrieval Augmented Generation.** 2024. 15 p. Preprint (arXiv:2402.05131). Disponível em: <https://arxiv.org/abs/2402.05131>. Acesso em: 26 nov. 2025.

KAYALI, B. *et al.* **Investigation of student experiences with ChatGPT-supported online learning applications in higher education.** Australasian Journal of Educational Technology, v. 39, n. 5, p. 20–39, 2023. Disponível em: <https://doi.org/10.14742/ajet.8915>. Acesso em: 26 nov. 2025.

KHAIRULLAH, S. A. *et al.* **Implementing artificial intelligence in academic and administrative processes through responsible strategic leadership in the higher education institutions.** Frontiers in Education, v. 10, 2025. Disponível em: <https://www.frontiersin.org/journals/education/articles/10.3389/feduc.2025.1548104>. Acesso em: 27 nov. 2025.

LEE, Sophia Soomin; MOORE, Robert L. **Harnessing Generative AI (GenAI) for Automated Feedback in Higher Education: A Systematic Review**. Online Learning, Boston, MA, v. 28, n. 3, 2024. DOI: 10.24059/olj.v28i3.4593. Disponível em: <https://olj.onlinelearningconsortium.org/index.php/olj/article/view/4593>. Acesso em: 27 nov. 2025.

MALA, Chandana sree *et al.* **Hybrid Retrieval for Hallucination Mitigation in Large Language Models: A Comparative Analysis**. 2025. 19 p. Preprint (arXiv:2504.05324). Disponível em: <https://arxiv.org/abs/2504.05324>. Acesso em: 27 nov. 2025.

NEUMANN, A. T. *et al.* **An LLM-Driven Chatbot in Higher Education for Databases and Information Systems**. IEEE Transactions on Education, v. 68, n. 1, p. 103-116, fev. 2025. Disponível em: <https://doi.org/10.1109/TE.2024.3467912>. Acesso em: 26 nov. 2025.

ROJAS, Marco Antonio Torrez; MEDEIROS, Jucelio Kulmann de. **Avaliação da adequação de Instituto Federal à Lei Geral de Proteção de Dados Pessoais**. Revista Eletrônica Argentina-Brasil de Tecnologias da Informação e da Comunicação, [S. l.], v. 1, n. 13, p. 1-20, jan. 2021. DOI 10.5281/zenodo.4445239. Disponível em: <https://revistas.setrem.com.br/index.php/reabtlic/article/view/391>. Acesso em: 27 nov. 2025.

SÁIZ-MANZANARES, María Consuelo *et al.* **Perceived satisfaction of university students with the use of chatbots as a tool for self-regulated learning**. Heliyon, v. 9, n. 1, e12843, 2023. DOI 10.1016/j.heliyon.2023.e12843. Disponível em: <https://doi.org/10.1016/j.heliyon.2023.e12843>. Acesso em: 27 nov. 2025.

SONG, M. **Enhancing RAG Performance by Representing Hierarchical Nodes in Headers for Tabular Data**. IEEE Access, v. 13, p. 85072-85083, 2025. DOI 10.1109/ACCESS.2025.3569872. Disponível em: <https://doi.org/10.1109/ACCESS.2025.3569872>. Acesso em: 26 nov. 2025.

SWACHA, J.; GRACEL, M. **Retrieval-Augmented Generation (RAG) Chatbots for Education: A Survey of Applications**. Applied Sciences, v. 15, n. 8, 4234, 2025. DOI 10.3390/app15084234. Disponível em: <https://doi.org/10.3390/app15084234>. Acesso em: 26 nov. 2025.

WANG, Xiaohua *et al.* **Searching for Best Practices in Retrieval-Augmented Generation.** 2024. 22 p. Preprint (arXiv:2407.01219). Disponível em: <https://arxiv.org/abs/2407.01219>. Acesso em: 27 nov. 2025.

WEI, Qikai *et al.* **QCG-Rerank:** Chunks Graph Rerank with Query Expansion in Retrieval-Augmented LLMs for Tourism Domain. 2024. 23 p. Preprint (arXiv:2411.08724). Disponível em: <https://arxiv.org/abs/2411.08724>. Acesso em: 27 nov. 2025.

YAN, Lixiang *et al.* **Practical and Ethical Challenges of Large Language Models in Education:** A Systematic Scoping Review. 2023. 23 p. Preprint (arXiv:2303.13379). Disponível em: <https://arxiv.org/abs/2303.13379>. Acesso em: 27 nov. 2025.

YANG, Yuming *et al.* **Benchmarking Multimodal RAG through a Chart-based Document Question-Answering Generation Framework.** 2025. 36 p. Preprint (arXiv:2502.14864). Disponível em: <https://arxiv.org/abs/2502.14864>. Acesso em: 27 nov. 2025.

ZHALNINA, O. L *et al.* Digitalization of Teaching: The Role of Artificial Intelligence in Higher Education. *Obščestvo: sociologija, psihologija, pedagogika*, [s. l.], n. 8, 2024.

ZOLFAGHARI, Vahid *et al.* **Adopting RAG for LLM-Aided Future Vehicle Design.** 2024. 6 p. Preprint (arXiv:2409.11960). Disponível em: <https://arxiv.org/abs/2409.11960>. Acesso em: 27 nov. 2025.