Acesso Inteligente à Informação Acadêmica: Assistente Institucional Baseado em Recuperação Documental e Inteligência Artificial

Jair Augusto Dério Boeing[1], **Vitor Hugo de Oliveira Morgante**[1], **Késsia Marchi**[1], **Marcelo Lopes Rosa**[1], **Rafael Henrique Dalegrave Zottesso**[1]

[1]Instituto Federal do Paraná – Campus Paranavaí – Paranavaí – PR – Brasil

{20233012578, 20233035895}@estudantes.ifpr.edu.br

{kessia.marchi, marcelo.rosa, rafael.zottesso}@ifpr.edu.br

Abstract. *This paper presents the development and preliminary evaluation of an institutional assistant based on document retrieval and generative artificial intelligence to support access to academic-administrative information. The prototype was developed for IFPR – Paranavaí Campus using an institutional document base and the Gemini API for response generation. The research follows a Design Science Research approach. A controlled evaluation was conducted with 21 students using computers with the same configuration and Internet infrastructure. The results indicated positive perceptions regarding response speed, objectivity, clarity, and access to information, although inconsistencies were observed.*

Resumo. *Este trabalho apresenta o desenvolvimento e a avaliação preliminar de um assistente institucional baseado em recuperação documental e inteligência artificial generativa para apoiar o acesso a informações acadêmico-administrativas. O protótipo foi desenvolvido para o IFPR – Campus Paranavaí utilizando uma base documental institucional e a API do Gemini para a geração das respostas. A pesquisa segue a abordagem Design Science Research. Foi realizada uma avaliação controlada com 21 estudantes, utilizando computadores com mesma configuração e infraestrutura de Internet. Os resultados indicaram percepções positivas quanto à rapidez, objetividade, clareza e acesso às informações, embora tenham sido observadas inconsistências.*

1. Introdução

A evolução das tecnologias digitais tem provocado mudanças na forma como instituições educacionais organizam seus processos de comunicação e acesso à informação [Almeida 2024, Soares 2025, Bordin et al. 2025]. Nesse contexto, documentos institucionais, calendários, regulamentos e orientações constituem importantes fontes de informação para estudantes. No Instituto Federal do Paraná (IFPR) – Campus Paranavaí, informações acadêmico-administrativas são disponibilizadas por meio de documentos institucionais. A motivação deste trabalho surgiu da necessidade observada no contexto do projeto de pesquisa de disponibilizar essas informações de maneira mais acessível, especialmente diante de dúvidas que exigem a localização de conteúdos específicos. Essa necessidade é tratada como motivação para o desenvolvimento do artefato

[Dresch et al. 2015], e não como resultado de um levantamento formal sobre a comunidade acadêmica.

Sistemas conversacionais podem oferecer uma alternativa de interação mais direta, permitindo que o estudante formule dúvidas em linguagem natural [Estrela et al. 2024, Bordin et al. 2025]. Entretanto, modelos generativos podem produzir respostas sem suporte documental, fenômeno conhecido na literatura como alucinação [Pereira and Aquino 2026]. Dessa forma, torna-se necessário associar a geração de linguagem à recuperação de informações provenientes de documentos previamente selecionados [Wensing et al. 2024, Medeiros and Oliveira 2025]. Uma abordagem utilizada para esse propósito é a Geração Aumentada por Recuperação, do inglês *Retrieval-Augmented Generation* (RAG), que combina mecanismos de recuperação com modelos generativos, utilizando informações recuperadas de uma fonte externa como contexto para a geração da resposta [Wensing et al. 2024, Medeiros and Oliveira 2025, Figueiredo et al. 2026].

Diante desse cenário, este trabalho tem como objetivo desenvolver e avaliar preliminarmente um protótipo de assistente institucional baseado em documentos do IFPR – Campus Paranavaí, utilizando recuperação da informação e inteligência artificial generativa. A avaliação considera clareza, objetividade, coerência com os documentos, tempo de resposta, facilidade de interação, capacidade de resposta e percepção dos participantes. Como contribuição, o trabalho apresenta um artefato aplicado ao contexto acadêmico-administrativo, descrevendo sua arquitetura, recuperação documental e geração das respostas, além de uma avaliação preliminar com estudantes.

2. Fundamentação Teórica

2.1. Chatbots e atendimento educacional

Chatbots são sistemas computacionais capazes de interagir com usuários por meio de linguagem natural, apoiando estudantes na obtenção de informações e na realização de consultas relacionadas à instituição [Estrela et al. 2024, Bordin et al. 2025]. No presente trabalho, o chatbot é utilizado como mecanismo complementar de acesso à informação institucional. Seu objetivo não é substituir servidores ou setores responsáveis pelo atendimento, mas apoiar a consulta a informações disponíveis em documentos oficiais. Trabalhos relacionados à aplicação de chatbots em ambientes acadêmicos demonstram a utilização desses sistemas como mecanismos de apoio à recuperação de informações e ao atendimento discente [Almeida 2024, Soares 2025, Bordin et al. 2025].

2.2. Recuperação da Informação

A Recuperação da Informação está relacionada à identificação e localização de informações relevantes em coleções de documentos. Ferneda [Ferneda 2012] apresenta conceitos relacionados à representação, indexação e recuperação de informações a partir de consultas realizadas pelos usuários. No contexto deste trabalho, a recuperação localiza trechos dos documentos institucionais relacionados à pergunta do estudante, funcionando como intermediária entre a consulta e a geração da resposta [Wensing et al. 2024]. Seu papel é central, pois o modelo generativo depende do contexto recuperado, podendo produzir respostas inadequadas quando informações relevantes não são localizadas ou selecionadas corretamente [Medeiros and Oliveira 2025, Figueiredo et al. 2026].

2.3. Inteligência Artificial Generativa

Modelos de inteligência artificial generativa podem produzir textos em linguagem natural a partir de instruções e informações fornecidas como entrada, permitindo transformar informações documentais em respostas próximas da linguagem utilizada pelos estudantes [Medeiros and Oliveira 2025, Bordin et al. 2025]. Esses modelos são treinados para identificar padrões em grandes volumes de dados e gerar novos conteúdos a partir do contexto recebido [Medeiros and Oliveira 2025, Pereira and Aquino 2026]. No entanto, a capacidade de produzir textos coerentes não garante que todas as informações geradas estejam corretas ou presentes nas fontes utilizadas (alucinação), aspecto relevante em aplicações institucionais [Wensing et al. 2024, Pereira and Aquino 2026].

Por esse motivo, o uso de modelos generativos em sistemas de recuperação pode ser associado a mecanismos que forneçam fontes externas de contexto [Wensing et al. 2024, Figueiredo et al. 2026]. Na abordagem utilizada, o conteúdo recuperado dos documentos institucionais é apresentado ao modelo juntamente com a pergunta do estudante, buscando reduzir a dependência de informações externas ao domínio delimitado [Medeiros and Oliveira 2025, Bordin et al. 2025]. No presente trabalho, a geração das respostas é realizada por meio da API do Gemini, utilizando o modelo Gemini 3.8 Flash (`gemini-3.8-flash`) durante os testes realizados em setembro de 2026, em integração similar à adotada por Bordin et al. [Bordin et al. 2025]. Essa separação entre recuperação documental e geração permite observar limitações específicas: quando informações relevantes não são recuperadas, o modelo recebe contexto insuficiente para responder adequadamente [Wensing et al. 2024, Figueiredo et al. 2026].

2.4. Geração Aumentada por Recuperação (RAG)

A Geração Aumentada por Recuperação, ou *Retrieval-Augmented Generation* (RAG), combina mecanismos de recuperação de informações com modelos generativos. Wensing et al. [Wensing et al. 2024], Medeiros e Oliveira [Medeiros and Oliveira 2025] e Figueiredo et al. [Figueiredo et al. 2026] descrevem essa arquitetura na qual o modelo gerador utiliza informações recuperadas de uma memória documental externa como contexto para a geração:

$$\textit{Pergunta} \rightarrow \textit{Recuperação} \rightarrow \textit{Contexto} \rightarrow \textit{Geração} \rightarrow \textit{Resposta}$$

No protótipo desenvolvido, essa abordagem é aplicada a uma coleção delimitada de documentos institucionais. A separação entre recuperação e geração permite analisar as limitações de cada etapa [Wensing et al. 2024, Figueiredo et al. 2026], visto que uma resposta inadequada pode decorrer tanto da recuperação de um contexto incorreto quanto da interpretação desse contexto pelo modelo.

2.5. Confiabilidade, rastreabilidade e proteção de dados

Em sistemas destinados ao atendimento acadêmico-administrativo, a confiabilidade das informações é relevante, pois uma resposta incorreta sobre procedimentos, datas ou regulamentos pode levar o estudante a interpretar inadequadamente uma norma institucional [Soares 2025, Bordin et al. 2025]. Por esse motivo, o protótipo utiliza uma base documental previamente selecionada, construindo a resposta a partir dos conteúdos recuperados [Wensing et al. 2024]. Além da confiabilidade, o tratamento de informações deve

considerar a Lei Geral de Proteção de Dados Pessoais (LGPD – Lei nº 13.709/2018), que estabelece princípios e regras para o tratamento de dados pessoais no Brasil [Brasil 2018]. No contexto deste trabalho, a avaliação foi realizada sem a coleta de informações destinadas à identificação individual dos participantes.

2.6. Trabalhos Correlatos

Soares [Soares 2025] desenvolveu uma pesquisa voltada à construção de uma base de conhecimento para recuperação e acesso à informação acadêmica para subsidiar um chatbot na Escola de Ciência da Informação da Universidade Federal de Minas Gerais (ECI/UFMG), aproximando-se deste estudo pelo uso de informações acadêmicas estruturadas. Almeida [Almeida 2024] investigou a possibilidade de implementação de um chatbot para atendimento ao discente na Universidade Federal de Campina Grande (UFCG), guardando relação por tratar do atendimento virtual estudantil.

Artigos brasileiros recentes também exploram chatbots e arquiteturas RAG no ensino público. Estrela et al. [Estrela et al. 2024] desenvolveram e avaliaram com estudantes o chatbot educacional *CosmoBot* utilizando escala Likert. Medeiros e Oliveira [Medeiros and Oliveira 2025] compararam modelos de *embeddings* e LLMs para RAG em língua portuguesa no Instituto Federal do Espírito Santo (IFES), enquanto Bordin et al. [Bordin et al. 2025] desenvolveram o *UniChat*, um assistente universitário baseado em RAG e na API do Gemini na Universidade Federal de Santa Maria (UFSM). O presente estudo diferencia-se ao combinar uma base documental institucional delimitada, recuperação de informações e geração via API do Gemini (modelo Gemini 3.8 Flash), realizando uma avaliação preliminar controlada diretamente com estudantes do ensino técnico integrado do IFPR – Campus Paranavaí.

3. Materiais e Métodos

3.1. Caracterização da pesquisa e Design Science Research

A pesquisa caracteriza-se como aplicada, pois envolve o desenvolvimento de um artefato tecnológico destinado a um problema prático relacionado ao acesso às informações acadêmico-administrativas, possuindo caráter exploratório quanto aos objetivos [Wazlawick 2020]. A abordagem combina aspectos quantitativos e qualitativos [Wazlawick 2020]: os quantitativos relacionam-se às respostas obtidas por meio de escala Likert [DalMoro and Vieira 2013] e à classificação das respostas produzidas pelo chatbot, enquanto os qualitativos referem-se às respostas abertas sobre percepções, dificuldades e sugestões de melhoria.

O desenvolvimento do protótipo foi orientado pela abordagem *Design Science Research* (DSR), voltada à construção e avaliação de artefatos destinados à solução de problemas identificados em determinado contexto [Dresch et al. 2015]. Conforme sistematizado por Dresch, Lacerda e Antunes Júnior [Dresch et al. 2015], o processo foi organizado nas etapas de definição do problema, levantamento e organização dos documentos, desenvolvimento do protótipo, implementação da recuperação e geração das respostas e avaliação preliminar com usuários.

3.2. Contexto, delimitação e construção da base documental

O estudo foi delimitado ao contexto do IFPR – Campus Paranavaí, sem a utilização de dados acadêmicos individualizados dos participantes e com domínio restrito às informações

disponíveis nos documentos selecionados. A concepção da base documental partiu da seleção de documentos institucionais relevantes para dúvidas acadêmico-administrativas (calendários, regulamentos e normas), preparados para localização de trechos pelo mecanismo de busca. Como estratégia metodológica de controle do domínio durante os testes com os estudantes, a base documental indexada no protótipo foi composta especificamente pelo Calendário Acadêmico de 2026 do IFPR – Campus Paranavaí, garantindo um cenário controlado no qual era possível identificar com exatidão quais datas e eventos deveriam servir de referência para as respostas e avaliar o bloqueio de perguntas externas ao calendário.Embora a arquitetura tenha sido concebida para permitir a incorporação de diferentes documentos institucionais, a avaliação apresentada neste trabalho foi realizada exclusivamente com o Calendário Acadêmico de 2026. Portanto, os resultados devem ser interpretados considerando esse domínio documental restrito, não sendo possível extrapolá-los diretamente para consultas sobre regulamentos ou outros procedimentos acadêmico-administrativos.

3.3. Critérios de avaliação

Os critérios de avaliação foram elaborados pelos autores com base em estudos nacionais sobre avaliação de chatbots no ensino [Estrela et al. 2024, Almeida 2024, Soares 2025, Bordin et al. 2025], contemplando doze dimensões: (1) clareza das respostas; (2) utilidade para esclarecer dúvidas; (3) relevância das informações; (4) coerência com os documentos institucionais; (5) objetividade das respostas; (6) facilidade de formular perguntas; (7) facilidade de compreender as respostas; (8) tempo necessário para receber as respostas; (9) percepção de facilidade de acesso às informações; (10) satisfação com a utilização; (11) intenção de utilizar novamente o sistema; e (12) capacidade de responder à pergunta realizada.

Os critérios foram operacionalizados em formulário eletrônico dividido em quatro partes: (i) perfil (vínculo de estudante do campus, curso e série); (ii) 11 afirmações em escala Likert de 5 pontos (de 1 – “Discordo totalmente” a 5 – “Concordo totalmente”) [DalMoro and Vieira 2013], cujos enunciados constam na Tabela 1; (iii) 4 questões objetivas sobre a consulta principal realizada pelo estudante: “A resposta apresentada pelo chatbot respondeu à pergunta realizada?” (Sim, completamente / Parcialmente / Não), “A resposta estava de acordo com o documento institucional?” (Sim / Não / Não foi possível verificar), e duas classificações em escala de 1 a 5 (clareza da resposta obtida e percepção geral sobre o chatbot); e (iv) 4 questões abertas (principal ponto positivo, principal problema ou dificuldade, sugestões de melhoria e descrição opcional de situação em que o chatbot não respondeu adequadamente).

3.4. Implementação do RAG

A implementação do mecanismo de Geração Aumentada por Recuperação foi organizada em duas etapas: indexação documental e recuperação durante a consulta. Na indexação, o texto do Calendário Acadêmico de 2026 do IFPR – Campus Paranavaí é dividido em trechos (*chunks*) de aproximadamente 450 *tokens*, com sobreposição de 50 *tokens*, preservando parágrafos e, quando necessário, frases. A contagem utiliza uma estimativa local de *tokens*. Cada trecho pode receber um vetor de representação (*embedding*) gerado pelo modelo `text-embedding-004`, utilizado na recuperação vetorial. O protótipo foi desenvolvido como aplicação web utilizando Python e JavaScript.

Durante a consulta, a pergunta é encaminhada a um roteador, que identifica a intenção, indica até três categorias documentais e extrai até cinco termos úteis à busca. Em seguida, a consulta é normalizada, com conversão para minúsculas, remoção de acentos e separação dos termos pela pontuação. Uma lista própria de palavras de parada (*stopwords*) em português é aplicada, descartando-se tokens com menos de três letras que não contenham dígitos.

A recuperação combina busca lexical e vetorial. A busca lexical considera a correspondência de tokens inteiros e variações simples de singular e plural. A consulta integral encontrada em um trecho acrescenta cinco pontos ao escore; cada termo encontrado acrescenta até quatro pontos no conteúdo e dois no título. Trechos associados às categorias indicadas pelo roteador recebem fator multiplicativo de 1,25 nesse escore. Na busca vetorial, calcula-se a similaridade de cosseno entre o *embedding* da consulta e os vetores dos trechos, com fator multiplicativo de 1,15 para as categorias indicadas. A combinação dessas estratégias permite considerar tanto a correspondência lexical quanto a proximidade semântica [Ferneda 2012, Medeiros and Oliveira 2025].

Cada mecanismo considera até 20 candidatos. As listas são combinadas pelo método *Reciprocal Rank Fusion* (RRF), cuja contribuição é calculada por $1/(60 + r)$ para cada lista, sendo r a posição do trecho no respectivo ranqueamento. A seleção final considera até cinco trechos, respeitando o orçamento de 2.400 *tokens* estimados para o contexto documental e descartando conteúdos repetidos. O primeiro trecho elegível é mantido mesmo quando excede individualmente esse orçamento. Na ausência de *embeddings*, a recuperação permanece exclusivamente lexical.

Por fim, o contexto reúne as instruções de sistema, os trechos selecionados e até quatro mensagens recentes do histórico, limitadas a 500 *tokens* estimados. Esse conteúdo é enviado à API do Gemini para geração da resposta. O roteador opera com temperatura zero e limite de 512 *tokens* de saída; o gerador utiliza temperatura 0,2, limite de 8.192 *tokens* de saída e nível de raciocínio baixo. As chamadas de geração e de *embedding* utilizam o SDK `google-genai`, com restrição de versão `google-genai>=1.47.0`. O identificador do modelo generativo corresponde à configuração utilizada nos testes. Após a geração, a resposta é apresentada ao estudante.

3.5. Avaliação com estudantes

A avaliação contou com 21 estudantes do IFPR – Campus Paranavaí, todos alunos do 4º ano do curso Técnico em Informática Integrado ao Ensino Médio, sem exceção, com faixa etária entre 17 e 19 anos. Os participantes foram convidados durante a aula de Desenvolvimento de Projetos, em atividade pedagógica supervisionada pelos docentes. Todos receberam esclarecimentos prévios sobre a pesquisa e manifestaram consentimento e assentimento livre e esclarecido para participação voluntária, sem coleta de dados de identificação individual em conformidade com a LGPD [Brasil 2018].

O experimento foi realizado em ambiente controlado, utilizando computadores com a mesma configuração e infraestrutura de Internet. Os participantes foram informados de que se tratava de um teste e que o chatbot possuía apenas contexto relacionado ao calendário acadêmico. Cada estudante pôde formular livremente suas perguntas quanto ao conteúdo e à forma de escrita, sem número mínimo ou máximo de consultas, recebendo um limite de 50.000 *tokens* durante a interação (não extrapolado por nenhum partici-

pante). Embora a liberdade de formulação aproxime o teste do uso natural, ela constitui uma variável relevante, pois diferentes participantes puderam consultar informações semelhantes usando construções linguísticas distintas. Após a interação, as respostas ao formulário foram analisadas de forma descritiva (frequências absolutas, percentuais sobre respostas válidas e médias) e qualitativa.

4. Desenvolvimento do Protótipo

4.1. Arquitetura da aplicação e fluxo de interação

Conforme apresentado na Figura 1, o protótipo foi estruturado em três componentes principais: interface de interação, mecanismo de recuperação documental e camada de geração de respostas. Na Figura 1, observa-se a integração entre os módulos: o estudante formula a pergunta na interface web do chatbot; a consulta passa pelo pré-processamento e busca textual na base documental; e os trechos selecionados compõem o contexto enviado junto à pergunta original para a camada generativa (API Gemini), cuja resposta em linguagem natural retorna à interface.

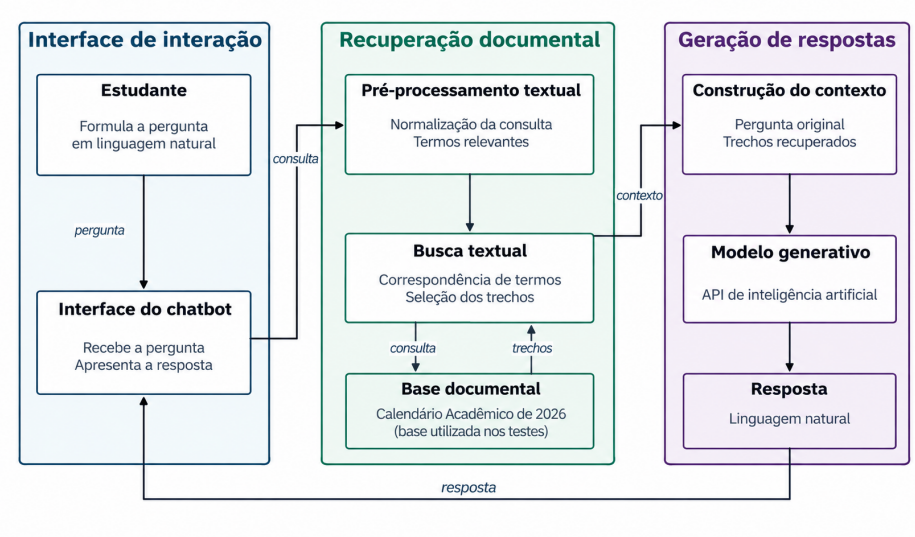


Figura 1. Arquitetura do assistente institucional baseado em recuperação documental e inteligência artificial generativa. Fonte: elaborado pelos autores.

Esse fluxo permite utilizar linguagem natural na interação sem abandonar a base documental definida, embora condicione a qualidade da resposta final à capacidade do mecanismo de recuperação de localizar os trechos relevantes [Wensing et al. 2024, Figueiredo et al. 2026].

4.2. Controle do domínio

O domínio do chatbot foi delimitado aos documentos institucionais selecionados para o projeto, buscando reduzir respostas sobre conteúdos externos ao sistema e evitar alucinações [Pereira and Aquino 2026]. Quando o contexto recuperado não apresenta informações suficientes, o comportamento configurado consiste em evitar a apresentação de informações sem suporte documental, orientando o usuário a procurar a Secretaria Acadêmica. Durante a avaliação, entretanto, foram observadas situações em que o sistema não conseguiu recuperar adequadamente o contexto necessário.

5. Resultados e Discussão

5.1. Resultados da avaliação

A avaliação contou com 21 respostas de estudantes. Os resultados quantitativos estão consolidados na Tabela 1, que reúne o total de respondentes em cada critério (N), a quantidade e o percentual de avaliações positivas (respostas "Concordo parcialmente" [4] e "Concordo totalmente" [5], ou notas 4 e 5) e a média obtida (escala de 1 a 5).

Tabela 1. Resultados quantitativos da avaliação pelos estudantes. Fonte: elaborada pelos autores.

Critério Avaliado (Escala Likert de 1 a 5)	N*	**Positivas – n (%)**	**Média**
Tempo necessário para receber as respostas adequado	21	18 (85,7%)	4,19
Objetividade das respostas às perguntas	21	17 (81,0%)	4,29
Respostas claras e fáceis de compreender	21	17 (81,0%)	4,05
Facilidade de compreender as respostas apresentadas	20	16 (80,0%)	4,20
Coerência com os documentos institucionais	21	16 (76,2%)	4,00
Relevância das informações para a pergunta realizada	20	15 (75,0%)	3,80
Facilidade de formular perguntas e interagir	21	15 (71,4%)	3,95
Facilidade de acesso às informações acadêmicas	21	15 (71,4%)	3,95
Intenção de utilizar novamente o chatbot	21	15 (71,4%)	3,81
Utilidade das respostas para esclarecer dúvidas	21	14 (66,7%)	3,81
Satisfação com a experiência de utilização	21	12 (57,1%)	3,62
Avaliação Direta por Nota (Escala de 1 a 5)	N	**Notas 4 e 5 – n (%)**	**Média**
Classificação da clareza da resposta obtida	21	15 (71,4%)	3,90
Percepção geral sobre o chatbot	21	14 (66,7%)	3,95

*Nota: Dois critérios apresentaram $N = 20$ devido à ausência de resposta de um participante na questão.

Conforme apresentado na Tabela 1, as percepções foram predominantemente positivas quanto à rapidez, objetividade, clareza e facilidade de compreensão (com aprovação entre 80,0% e 85,7% e médias superiores a 4,00), seguidas pela coerência documental, relevância, facilidade de interação, acesso à informação e intenção de uso futuro (entre 71,4% e 76,2%). Nos itens de facilidade de compreensão e relevância, houve uma abstenção em cada questão ($N = 20$), resultando nos percentuais exatos de 80,0% (16/20) e 75,0% (15/20). Por outro lado, a utilidade (66,7%) e a satisfação geral (57,1%; média 3,62) apresentaram resultados mais heterogêneos, contrastando com os atributos formais: embora as respostas fossem rápidas, claras e objetivas, a ocorrência de respostas incompletas ou falhas de busca reduziu a satisfação quando a dúvida não foi integralmente atendida.

5.2. Capacidade de resposta e coerência documental

Para a classificação da capacidade de resposta na consulta principal do estudante ($N = 21$), considerou-se "completa" aquela que atendeu integralmente à pergunta, "parcial" quando apresentou informações insuficientes e "não respondida" quando não atendeu à dúvida. Dos 21 participantes, 10 (47,6%) indicaram resposta completa, 8 (38,1%) parcial e 3 (14,3%) não respondida, totalizando 18 participantes (85,7%) com atendimento completo ou parcial. Na verificação da correspondência com os documentos institucionais, 17 participantes (81,0%) indicaram que a resposta estava de acordo com o documento, 3

(14,3%) informaram não ter sido possível verificar e 1 (4,8%) indicou discordância. Esses índices refletem a percepção dos usuários, não constituindo validação independente da correção factual.

5.3. Percepção dos participantes

As respostas abertas reforçaram os resultados da Tabela 1, destacando o acesso mais rápido às informações, a rapidez, a objetividade e a clareza. Entre os comentários registrados, destacam-se percepções como "acesso mais rápido a informação", "rapidez", "a ideia é muito boa e realmente ajudaria" e "ele facilita a retirada de dúvidas e responde com clareza". Esses relatos confirmam a recepção favorável à formulação de consultas em linguagem natural, embora representem a percepção subjetiva da experiência de uso, e não uma comparação experimental com outras formas de consulta.

5.4. Inconsistências observadas

Apesar das avaliações positivas, foram observadas limitações importantes no funcionamento do sistema. Alguns participantes relataram informações incorretas ou incompletas e situações em que informações presentes no calendário acadêmico (como início e fim do trimestre ou a data da SETIF) não foram recuperadas. Também foram observadas respostas que terminavam antes da conclusão da frase ("respostas quebradas ao meio", associadas ao limite de *tokens* de saída na geração) ou apresentavam contexto insuficiente. Por outro lado, em perguntas fora do escopo do calendário (como sobre abono de faltas por atestado de doação de sangue), o sistema informou corretamente a ausência da informação nos documentos disponíveis e orientou a procura pela Secretaria do campus.

Um dos principais problemas identificados foi a influência da forma como a pergunta era formulada. Nas consultas sobre os feriados de novembro, ao perguntar "Quais feriados tem em novembro?", o sistema não conseguiu identificar adequadamente a dúvida e orientou o usuário a consultar o material; já na reformulação "E de acordo com o calendário acadêmico, quais são os feriados de novembro?", o sistema recuperou o contexto adequado e apresentou corretamente as datas. Esse comportamento é crítico em uma arquitetura RAG, pois a geração depende dos trechos recuperados [Wensing et al. 2024, Medeiros and Oliveira 2025]: se a recuperação falha, o modelo generativo produz uma resposta incompleta ou inadequada mesmo quando a informação consta na base [Figueiredo et al. 2026].

5.5. Discussão

Os resultados evidenciam aspectos positivos e limitações que dialogam com a literatura nacional. Os ganhos percebidos em rapidez, objetividade e clareza (Tabela 1) convergem com os estudos de Soares [Soares 2025], Almeida [Almeida 2024], Estrela et al. [Estrela et al. 2024] e Bordin et al. [Bordin et al. 2025], que demonstram a boa aceitação discente a chatbots voltados à consulta ágil de informações acadêmicas.

Em contrapartida, a diferença entre a avaliação elevada das características formais da interação e a satisfação geral mais moderada (Tabela 1), somada à variação de respostas para perguntas semanticamente próximas (como no caso dos feriados de novembro), corrobora as análises de Wensing et al. [Wensing et al. 2024], Medeiros e Oliveira [Medeiros and Oliveira 2025] e Figueiredo et al. [Figueiredo et al. 2026]. Como a busca

utilizou correspondência léxica de termos [Ferneda 2012], pequenas alterações de vocabulário modificaram o contexto recuperado, impactando a satisfação quando a resposta saiu incompleta, ainda que o controle de domínio tenha contido alucinações em perguntas fora do escopo [Pereira and Aquino 2026]. Ressalta-se, ainda, que a percepção de facilidade de acesso reflete apenas a experiência declarada pelos usuários, pois não houve comparação experimental com a busca tradicional.

5.6. Limitações do estudo e implicações para o artefato

A avaliação realizada apresenta limitações que delimitam suas conclusões: (i) a amostra contou com 21 estudantes exclusivamente do 4º ano do curso Técnico em Informática (17 a 19 anos), não devendo ser generalizada para alunos de outros cursos, ingressantes ou servidores; (ii) o domínio de teste restringiu-se ao calendário acadêmico; (iii) não houve comparação experimental com métodos convencionais para medir redução objetiva no tempo de busca; e (iv) a livre formulação de consultas não contemplou um conjunto padronizado de perguntas nem métricas técnicas objetivas (como tempo de resposta em segundos, taxa de recuperação correta dos documentos e validação independente questão por questão).

Na perspectiva da *Design Science Research* [Dresch et al. 2015], esses resultados indicam requisitos diretos para os próximos ciclos de desenvolvimento do artefato: incorporar mecanismos de recuperação densa baseados em *embeddings* semânticos e busca híbrida [Medeiros and Oliveira 2025, Figueiredo et al. 2026], aprimorar a segmentação (*chunking*) dos documentos, ajustar os parâmetros de saída de *tokens* na API para evitar respostas truncadas e aperfeiçoar o controle explícito de insuficiência de contexto [Wensing et al. 2024, Pereira and Aquino 2026].

6. Conclusão

Este trabalho apresentou o desenvolvimento e a avaliação preliminar de um assistente institucional baseado em recuperação documental e inteligência artificial generativa (modelo Gemini 3.8 Flash) para apoiar o acesso a informações acadêmico-administrativas do IFPR – Campus Paranavaí. A avaliação controlada com 21 estudantes indicou percepções predominantemente positivas quanto ao tempo de resposta (85,7

Como contribuição, o trabalho apresenta um artefato funcional e uma avaliação empírica preliminar capaz de identificar potencialidades e limitações técnicas da solução. Como trabalhos futuros, pretende-se implementar recuperação semântica e híbrida, aprimorar a segmentação documental, expandir a base institucional, eliminar truncamentos na geração, aplicar um conjunto padronizado de perguntas com métricas técnicas objetivas, realizar comparações experimentais com a consulta convencional e ampliar a avaliação para estudantes de outros cursos, ingressantes e servidores.

Apêndice A – Instrumento de Avaliação Aplicado

Bloco 1 – Perfil: (1) Você é estudante do IFPR Campus Paranavaí? [Sim/Não]; (2) Curso; (3) Ano ou Série.
Bloco 2 – Escala Likert (1 = Discordo totalmente a 5 = Concordo totalmente): (1) As respostas fornecidas pelo chatbot foram claras e fáceis de compreender; (2) As respostas fornecidas pelo chatbot foram úteis para esclarecer minhas dúvidas; (3) O chatbot apresentou informações

relevantes para a pergunta realizada; (4) As respostas apresentadas estavam coerentes com as informações dos documentos institucionais; (5) O chatbot respondeu às perguntas de forma objetiva; (6) Foi fácil formular perguntas e interagir com o chatbot; (7) Foi fácil compreender as respostas apresentadas pelo chatbot; (8) O tempo necessário para receber as respostas foi adequado; (9) O chatbot facilitou o acesso às informações acadêmico-administrativas; (10) Eu me senti satisfeito com a experiência de utilização do chatbot; (11) Eu utilizaria novamente o chatbot para consultar informações acadêmico-administrativas.

Bloco 3 – Verificação Objetiva: (12) A resposta apresentada pelo chatbot respondeu à pergunta realizada? [Sim, completamente / Parcialmente / Não]; (13) A resposta estava de acordo com o documento institucional? [Sim / Não / Não foi possível verificar]; (14) Como você classificaria a clareza dessa resposta? [Escala 1 a 5]; (15) Qual foi sua percepção geral sobre o chatbot? [Escala 1 a 5].

Bloco 4 – Questões Abertas: (16) Qual foi o principal ponto positivo do chatbot?; (17) Qual foi o principal problema ou dificuldade encontrada durante o uso?; (18) O que você sugere para melhorar o chatbot?; (19) Caso queira, descreva alguma situação em que o chatbot não conseguiu responder adequadamente à sua pergunta.

Referências

Almeida, E. J. C. (2024). Otimização da comunicação institucional na UFCG: examinando a possibilidade de implementação de um chatbot para atendimento ao discente. Dissertação (mestrado profissional em administração pública), Universidade Federal de Campina Grande, Sousa.

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 XL Simpósio Brasileiro de Bancos de Dados (SBBD)*, Porto Alegre. SBC.

Brasil (2018). Lei nº 13.709, de 14 de agosto de 2018. Lei Geral de Proteção de Dados Pessoais (LGPD). Diário Oficial da União, Brasília, DF. Acesso em: 15 out. 2026.

DalMoro, M. and Vieira, K. M. (2013). Dilemas na construção de escalas tipo Likert: o número de itens e a disposição influenciam nos resultados? *Revista Gestão Organizacional*, 6(3):161–174.

Dresch, A., Lacerda, D. P., and Antunes Júnior, J. A. V. (2015). *Design Science Research: Método de Pesquisa para Avanço da Ciência e Tecnologia*. Bookman, Porto Alegre.

Estrela, I. R. B., Barbosa, K. B., and Brito, C. P. L. (2024). Ferramenta CosmoBot: Um chatbot de apoio a alunos em avaliação de algoritmos. In *Anais do XXXV Simpósio Brasileiro de Informática na Educação (SBIE)*, pages 380–391, Porto Alegre. SBC.

Ferneda, E. (2012). *Introdução aos Modelos Computacionais de Recuperação de Informação*. Ciência Moderna, Rio de Janeiro.

Figueiredo, P., Cardoso, B. O., Fontoura, J. R. d. A., and Santana, E. N. d. (2026). Retrieval-augmented generation (RAG) híbrido: arquiteturas, métodos de recuperação e desafios para sistemas baseados em large language models. *Contribuciones a las Ciencias Sociales*, 19(6):e24482.

Medeiros, L. S. F. and Oliveira, H. T. A. d. (2025). Comparação de modelos de embeddings e LLMs para geração aumentada por recuperação em português. In *Anais do LII Seminário Integrado de Software e Hardware (SEMISH)*, pages 429–440, Porto Alegre. SBC.

Pereira, A. S. and Aquino, S. A. B. d. M. (2026). Alucinações em modelos de linguagem de grande escala: Uma revisão sistemática da literatura. In *Anais do LIII Seminário Integrado de Software e Hardware (SEMISH)*, Porto Alegre. SBC.

Soares, J. R. (2025). Base de conhecimento para recuperação e acesso à informação acadêmica: provimento para um chatbot na ECI UFMG. Dissertação (mestrado em gestão e organização do conhecimento), Universidade Federal de Minas Gerais, Belo Horizonte.

Wazlawick, R. S. (2020). *Metodologia de Pesquisa para Ciência da Computação*. LTC, Rio de Janeiro, 3 edition.

Wensing, I. M., Joppi Filho, N. L., Dorneles, C., and Albuquerque, A. M. d. (2024). Avaliação de aplicações de geração aumentada de recuperação por meio de feedback implícito. In *Anais Estendidos do XXXIX Simpósio Brasileiro de Bancos de Dados (SBBD)*, Porto Alegre. SBC.