



INSTITUTO FEDERAL GOIANO
CAMPUS URUTAÍ
NÚCLEO DE INFORMÁTICA
CURSO DE GESTÃO DA TECNOLOGIA DA INFORMAÇÃO

VIRGINIA ELYZABETH SPRICIGO

CHATBOT RAG PERSONALIZADO DE DADOS EM MODELOS DE LINGUAGEM NATURAL GENERATIVOS

Urutaí, 14 de março de 2026

INSTITUTO FEDERAL GOIANO
NÚCLEO DE INFORMÁTICA
GESTÃO DA TECNOLOGIA DA INFORMAÇÃO

VIRGINIA ELYZABETH SPRICIGO

CHATBOT RAG PERSONALIZADO DE DADOS EM MODELOS DE LINGUAGEM NATURAL GENERATIVOS

Monografia apresentada ao Núcleo de Informática, curso de Gestão da Tecnologia da Informação, do Instituto Federal Goiano, como parte das exigências para obtenção do título de Tecnólogo em Gestão da Tecnologia da Informação.

Orientador(a):
Prof. Dr. GABRIEL DA SILVA VIEIRA

Urutaí, 14 de março de 2026

**Ficha de identificação da obra elaborada pelo autor, através do
Programa de Geração Automática do Sistema Integrado de Bibliotecas do IF Goiano - SIBi**

S769 Spricigo, Virginia Elyzabeth
CHATBOT RAG PERSONALIZADO DEDADOS EM
MODELOS DE LINGUAGEM NATURAL GENERATIVOS /
Virginia Elyzabeth Spricigo. Urutaí 2026.
32f. il.
Orientador: Prof. Dr. Gabriel da Silva Vieira.
Tcc (Especialista) - Instituto Federal Goiano, curso de 0121103 -
Tecnologia em Gestão da Tecnologia da Informação - Urutaí
(Campus Urutaí).
1. Chatbot. 2. Geração Aumentada por Recuperação. 3. Grandes
Modelos de Linguagem. 4. Google Colab. 5. LlamaIndex. I.
Título.

TERMO DE CIÊNCIA E DE AUTORIZAÇÃO

PARA DISPONIBILIZAR PRODUÇÕES TÉCNICO-CIENTÍFICAS

NO REPOSITÓRIO INSTITUCIONAL DO IF GOIANO

Com base no disposto na Lei Federal nº 9.610, de 19 de fevereiro de 1998, AUTORIZO o Instituto Federal de Educação, Ciência e Tecnologia Goiano a disponibilizar gratuitamente o documento em formato digital no Repositório Institucional do IF Goiano (RIIF Goiano), sem ressarcimento de direitos autorais, conforme permissão assinada abaixo, para fins de leitura, download e impressão, a título de divulgação da produção técnico-científica no IF Goiano.

IDENTIFICAÇÃO DA PRODUÇÃO TÉCNICO-CIENTÍFICA

Tese (doutorado)

Dissertação (mestrado)

Monografia (especialização)

TCC (graduação)

Artigo científico

Capítulo de livro

Livro

Trabalho apresentado em evento

Produto técnico e educacional - Tipo:

Nome completo do autor:

Matrícula:

Título do trabalho:

RESTRIÇÕES DE ACESSO AO DOCUMENTO

Documento confidencial: Não Sim, justifique:

Informe a data que poderá ser disponibilizado no RIIF Goiano: / /

O documento está sujeito a registro de patente? Sim Não

O documento pode vir a ser publicado como livro? Sim Não

DECLARAÇÃO DE DISTRIBUIÇÃO NÃO-EXCLUSIVA

O(a) referido(a) autor(a) declara:

- Que o documento é seu trabalho original, detém os direitos autorais da produção técnico-científica e não infringe os direitos de qualquer outra pessoa ou entidade;
- Que obteve autorização de quaisquer materiais inclusos no documento do qual não detém os direitos de autoria, para conceder ao Instituto Federal de Educação, Ciência e Tecnologia Goiano os direitos requeridos e que este material cujos direitos autorais são de terceiros, estão claramente identificados e reconhecidos no texto ou conteúdo do documento entregue;
- Que cumpriu quaisquer obrigações exigidas por contrato ou acordo, caso o documento entregue seja baseado em trabalho financiado ou apoiado por outra instituição que não o Instituto Federal de Educação, Ciência e Tecnologia Goiano.

Local

Data

Documento assinado digitalmente

VIRGINIA ELYZABETH SPRICIGO

Data: 14/03/2026 12:31:46-0300

Verifique em <https://validar.iti.gov.br>

Assinatura do autor e/ou detentor dos direitos autorais

Ciente e de acordo:

Assinatura do(a) orientador(a)

Documento assinado digitalmente

GABRIEL DA SILVA VIEIRA

Data: 13/03/2026 22:08:15-0300

Verifique em <https://validar.iti.gov.br>



SERVIÇO PÚBLICO FEDERAL
MINISTÉRIO DA EDUCAÇÃO
SECRETARIA DE EDUCAÇÃO PROFISSIONAL E TECNOLÓGICA
INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA GOIANO

Ata nº 5/2026 - CCBSI-URT/GE-UR/DE-UR/CMPURT/IFGOIANO

CURSO SUPERIOR DE GESTÃO DA TECNOLOGIA INFORMAÇÃO
ATA DE APRESENTAÇÃO DE TRABALHO DE CURSO

Aos treze dias do mês de fevereiro de dois mil e vinte e seis, reuniram-se os professores: Gabriel da Silva Vieira, Rachel Lopes Carcute, e Luciana Teles dos Santos Mesquita de Sousa nas dependências do Instituto Federal Goiano - Campus Urutaí, para avaliar o Trabalho de Curso da acadêmica: Virginia Elyzabeth Spricigo, como requisito necessário para a conclusão do Curso Superior de Gestão da Tecnologia da Informação desta Instituição. O presente TC tem como título: **Chatbot Rag Personalizado de Dados em Modelos de Linguagem Natural Generativos**, foi orientado por Gabriel da Silva Vieira.

Após análise, foram dadas as seguintes notas:

Professores	Aluno / Notas
	Virginia Elyzabeth Spricigo
1. Gabriel da Silva Vieira	9.0
2. Rachel Lopes Carcute	9.0
3. Luciana Teles dos Santos Mesquita de Sousa	9.5
MÉDIA FINAL:	9.1

OBSERVAÇÕES:

Por ser verdade firmamos a presente:

(Assinado Eletronicamente)

Gabriel da Silva Vieira

Orientador

(Assinado Eletronicamente)

Rachel Lopes Carcute

Membro

(Assinado Eletronicamente)

Luciana Teles dos Santos Mesquita de Sousa

Membro

Documento assinado eletronicamente por:

- **Gabriel da Silva Vieira, PROFESSOR ENS BASICO TECN TECNOLOGICO** , em 13/02/2026 20:06:51.
- **Rachel Lopes Carcute, PROFESSOR ENS BASICO TECN TECNOLOGICO** , em 13/02/2026 20:10:56.
- **Luciana Teles dos Santos Mesquita de Sousa, PROF ENS BAS TEC TECNOLOGICO-SUBSTITUTO** , em 13/02/2026 20:17:36.

Este documento foi emitido pelo SUAP em 13/02/2026. Para comprovar sua autenticidade, faça a leitura do QRCode ao lado ou acesse <https://suap.ifgoiano.edu.br/autenticar-documento/> e forneça os dados abaixo:

Código Verificador: 789423

Código de Autenticação: 9139c5a100



INSTITUTO FEDERAL GOIANO

Campus Urutaí

Rodovia Geraldo Silva Nascimento, Km 2.5, SN, Zona Rural, URUTAÍ / GO, CEP 75790-000

(64) 3465-1900

Este trabalho apresenta o desenvolvimento de um *chatbot* baseado em Geração Aumentada por Recuperação (RAG), adaptado para processar os dados dos editais do IF Goiano Campus Urutaí, especificamente os de pesquisa, regulamento de graduação e assistência estudantil. O objetivo do sistema é otimizar o tempo de consulta de alunos e professores e reduzir a carga de trabalho administrativo das secretarias e coordenadorias. A metodologia de criação foi dividida em cinco fases: coleta e preparação de dados, criação da base de conhecimento com *embeddings*, desenvolvimento do sistema RAG, construção da interface do usuário (*chatbot*) e, por fim, teste e avaliação do sistema. Após diversos testes e melhorias no banco de dados, obtiveram-se resultados positivos, demonstrando a viabilidade da solução, embora com potencial identificado para futuras melhorias e implementações.

Palavras-chave:

Chatbot. Geração Aumentada por Recuperação. Grandes Modelos de Linguagem.

This work presents the development of a chatbot based on Retrieval-Augmented Generation (RAG), adapted to process data from the official announcements of IF Goiano Campus Urutaí, specifically those related to research, graduation regulations, and student assistance. The system aims to optimize query time for students and professors and reduce the administrative workload of secretariats and coordinations. The development methodology was divided into five phases: data collection and preparation, creation of the knowledge base with embeddings, RAG system development, construction of the user interface (chatbot), and finally, system testing and evaluation. After several tests and database improvements, positive results were achieved, demonstrating the solution's viability, although potential for future improvements and implementations was also identified.

Key-words:

Chatbot. Retrieval-Augmented Generation. Large Language Models.

1	INTRODUÇÃO	4
1.1	Justificativa	5
1.2	Objetivos	5
1.2.1	Objetivo Geral	5
1.2.2	Objetivos Específicos	5
1.3	Estrutura do Trabalho	6
2	FUNDAMENTAÇÃO TEÓRICA	7
2.1	Grandes Modelos de Linguagem	7
2.2	Processamento de Linguagem Natural (PLN)	8
2.3	Chatbot	9
2.4	Geração Aumentada por Recuperação	10
3	METODOLOGIA	12
3.1	Coleta e Preparação de Dados	12
3.2	Criação da Base de Conhecimento	12
3.3	Desenvolvimento do Sistema RAG	13
3.4	Construção da Interface do Usuário	13
3.5	Teste e Avaliação	13
3.6	Ferramentas e recursos utilizados	14
4	RESULTADOS E DISCUSSÃO	15
4.1	Apresentação da Interface do Usuário	15
4.2	Análise de Resultados	16
4.3	Discussão: Limitações e o Processo de Refinamento	22
5	CONCLUSÃO	25
	CONCLUSÃO	25

CAPÍTULO 1

INTRODUÇÃO

A Geração Aumentada por Recuperação (RAG) é uma técnica avançada de inteligência artificial que capacita Grandes Modelos de Linguagem (LLMs) a acessar e integrar informações externas de uma base de conhecimento específica. Em vez de depender apenas do conhecimento internalizado durante seu treinamento, o RAG permite que os LLMs recuperem dados relevantes em tempo real, garantindo que as respostas geradas sejam não só coerentes, mas também factualmente precisas e contextualizadas (MANGLA, 2024).

Nos últimos anos, a aplicação de RAG tem se destacado em pesquisas internacionais voltadas ao aprimoramento de sistemas de pergunta e resposta, assistentes virtuais e ferramentas de apoio à tomada de decisão baseadas em conhecimento especializado. Diversos estudos publicados majoritariamente em língua inglesa investigam o uso dessa abordagem para integrar modelos de linguagem com bases documentais específicas, permitindo consultas mais precisas em ambientes corporativos, acadêmicos e institucionais. Esses trabalhos demonstram que a combinação entre modelos de linguagem e mecanismos de recuperação de informação pode reduzir alucinações dos modelos e melhorar significativamente a confiabilidade das respostas geradas.

O volume crescente de informações e a complexidade inerente a documentos oficiais, como editais universitários, representam um desafio significativo para a comunidade acadêmica. Alunos e professores dedicam considerável tempo na busca por dados específicos em textos extensos e densos, o que pode levar à perda de oportunidades e à subutilização de recursos institucionais essenciais. Nesse contexto, os editais representam documentos oficiais e técnicos cruciais em instituições de ensino, como o IF Goiano Campus Urutaí, contendo infor-

mações vitais sobre pesquisas científicas, regulamentos de graduação e assistência estudantil. A extensão e complexidade desses documentos dificultam o acesso rápido e preciso à informação, resultando em perda de tempo para alunos e professores e em ineficiência que culmina em gargalos informacionais.

1.1 Justificativa

A economia de tempo para alunos e professores na busca por informações, a redução da carga de trabalho administrativo para as secretarias e coordenadorias que atualmente respondem a essas dúvidas manualmente, e o potencial de aumentar a participação em programas e pesquisas, são benefícios diretos e tangíveis. Ao fornecer acesso rápido e claro aos editais de pesquisa científica, regulamento de graduação e assistência estudantil, o sistema proposto não só promove a eficiência operacional, mas também fortalece o engajamento da comunidade acadêmica, permitindo que o foco seja direcionado para atividades de maior valor agregado, como o ensino, a pesquisa e a extensão.

1.2 Objetivos

1.2.1 Objetivo Geral

Apresentar o desenvolvimento de um *Chatbot RAG* com IA (Inteligência Artificial) para auxiliar no entendimento, leitura e busca de informação de editais do IF Goiano Campus Urutaí.

1.2.2 Objetivos Específicos

- a) Analisar e coletar os editais de pesquisa, regulamento de graduação e assistência estudantil do IF Goiano Campus Urutaí em formatos digitais.
- b) Pré-processar e indexar o conteúdo dos editais, utilizando técnicas de *embedding* para criar uma base de conhecimento otimizada.
- c) Implementar e integrar um sistema de Geração Aumentada por Recuperação (RAG) com um Modelo de Linguagem Grande (LLM) para respostas contextuais.
- d) Desenvolver uma interface de usuário (*chatbot*) intuitiva para a interação entre alunos, professores e o sistema de consulta de editais.
- e) Avaliar a precisão e a usabilidade do *chatbot* por meio de testes e feedback de usuários do IF Goiano Campus Urutaí.

1.3 Estrutura do Trabalho

O presente trabalho está estruturado em cinco capítulos para detalhar o processo de concepção e desenvolvimento da solução. O capítulo [1](#) versa sobre a introdução desse trabalho. O capítulo [2](#) apresenta a Fundamentação Teórica, abordando os conceitos de LLMs, *Chatbots* e a arquitetura RAG. O Capítulo [3](#) descreve a Metodologia da pesquisa, detalhando as cinco fases de desenvolvimento do projeto, desde a coleta de dados até os testes. O Capítulo [4](#) apresenta e discute os resultados obtidos, incluindo a análise da interface e o desempenho do *chatbot*. Por fim, o Capítulo [5](#) apresenta as Considerações Finais, resumindo as conclusões do trabalho e sugerindo caminhos para pesquisas futuras.

CAPÍTULO 2

FUNDAMENTAÇÃO TEÓRICA

Esta seção aborda os principais conceitos técnicos que sustentam o desenvolvimento deste trabalho, com foco em Grandes Modelos de Linguagem(LLMs), Processamento de Linguagem Natural (PLN), *Chatbot* e Geração Aumentada por Recuperação(RAG).

2.1 Grandes Modelos de Linguagem

Para entender a tecnologia por trás do *chatbot* desenvolvido, é fundamental começar pelos Grandes Modelos de Linguagem (LLMs). Conforme afirma [Xiao Tong Zhu \(2025\)](#), tradução nossa), "os grandes modelos de linguagem originaram-se do processamento de linguagem natural, mas sem dúvida se tornaram um dos avanços tecnológicos mais revolucionários no campo da inteligência artificial nos últimos anos".

Uma das principais percepções associadas aos grandes modelos de linguagem é que o conhecimento sobre o mundo e sobre as línguas pode ser adquirido por meio de tarefas de modelagem de linguagem realizadas em larga escala. A partir dessa abordagem, torna-se possível desenvolver modelos mais abrangentes, capazes de lidar com diferentes tipos de problemas relacionados ao processamento de linguagem natural. Essa mudança trouxe impactos significativos nas metodologias de pesquisa da área, que passaram a substituir o treinamento de sistemas especializados desde o início, baseado em grandes volumes de dados rotulados, pelo uso de modelos previamente treinados em larga escala, os quais podem ser posteriormente ajustados e alinhados para aplicações específicas.

De acordo com [Varas et al. \(2023\)](#), os modelos de linguagem de grande escala (LLMs), classificados como uma categoria de inteligência artificial generativa, têm recebido grande aten-

ção nos últimos anos. Segundo os autores, essa relevância está relacionada à capacidade desses modelos de produzir textos semelhantes à escrita humana a partir de grandes volumes de dados utilizados em seu treinamento.

Ainda conforme [Varas et al. \(2023\)](#), esses modelos são treinados com extensas quantidades de textos provenientes de diferentes fontes, o que lhes permite identificar padrões linguísticos e gerar respostas contextualmente adequadas em diversos idiomas e domínios.

A partir dessas características, observa-se que os LLMs vêm sendo aplicados em diferentes contextos tecnológicos, incluindo assistentes virtuais, sistemas de perguntas e respostas e ferramentas de geração automatizada de conteúdo. Entre os exemplos mais conhecidos dessa tecnologia destacam-se os modelos GPT-3 e GPT-4, desenvolvidos pela OpenAI.

Eles também apresentam um desempenho robusto em uma variedade de atividades de Processamento de Linguagem Natural (PLN). Tipicamente, são usados para tarefas que envolvem geração de texto, sistemas de perguntas e respostas, tradução automática ou análise de sentimentos.

2.2 Processamento de Linguagem Natural (PLN)

O processamento da linguagem natural (PLN) trata computacionalmente os diversos aspectos comunicação humana, como sons, palavras, sentenças e discursos, considerando formatos e referências, estruturas e significados, contextos e usos. Em sentido bem amplo, podemos dizer que o PLN visa fazer o computador se comunicar em linguagem humana, nem sempre necessariamente em todos os níveis de entendimento e/ou geração de sons, palavras, sentenças e discursos.

PLN vem sendo estudada desde 1946 com objetivando que o computador possa se comunicar naturalmente através da linguagem humana. Ao longo do tempo ocorreu várias evoluções sobre o tema, houve a criação de várias ferramentas para desenvolvimento de aplicações de PLN, aplicações como análise fonética, análise morfológica, análise sintática e análise semântica. Análise fonética é o reconhecimento de sons que está presente nas palavras, utilizada frequentemente no processamento da voz humana para texto. Análise morfológica realiza a análise da estrutura, formação e classificações das palavras para que seja possível inferir a que se refere determinada palavra, determinando, por exemplo, se uma palavra é um substantivo ou adjetivo.

De acordo com [\(CARVALHO, 2003\)](#), análise sintática tem por objetivo identificar como

se relacionam as palavras entre si, e o seu papel na oração. Análise semântica analisa o significado das palavras e seu relacionamento na frase, para construir as sentenças, buscando identificar o contexto em que uma palavra está inserida.

De modo geral, conforme (CASELI et al., 2023), em PLN buscam-se soluções para problemas computacionais, ou seja, tarefas, sistemas, aplicações ou programas, que requerem o tratamento computacional de uma língua (português, inglês etc.), seja escrita (texto) ou falada (fala). Línguas como as de sinais também têm sido alvo de estudos da área. Cada modo tem suas especificidades. No caso da fala, as características que a distinguem da língua escrita são relacionadas a questões da produção (síntese) e recepção (reconhecimento) do som.

2.3 Chatbot

Uma das aplicações mais populares dos modelos de linguagem é a criação de *chatbots*. No contexto educacional, essas ferramentas vêm ganhando destaque devido ao seu potencial de apoio às atividades de ensino e aprendizagem.

De acordo com (Sousa et al., 2023), os *chatbots* são programas de computador desenvolvidos para simular conversas humanas e interagir com os usuários. Segundo os autores, no ambiente educacional essas ferramentas podem auxiliar os educadores ao responder dúvidas frequentes dos estudantes, fornecer *feedback* imediato e personalizado e até mesmo apoiar a elaboração de materiais educacionais. Além disso, seu uso pode contribuir para a redução da carga de trabalho docente e para a melhoria da eficiência das atividades de ensino.

Ainda conforme (Sousa et al., 2023), outra aplicação da inteligência artificial na atuação docente envolve a análise de dados educacionais. Por meio de algoritmos de aprendizado de máquina, torna-se possível coletar e analisar grandes volumes de dados relacionados ao desempenho dos alunos, às suas preferências de aprendizagem e ao seu nível de engajamento. Essas informações podem ser utilizadas para identificar padrões, realizar previsões e gerar *insights* que auxiliem os educadores na adaptação de suas práticas pedagógicas e na oferta de experiências de aprendizagem mais personalizadas.

Os *bots*, termo derivado da palavra *robot*, são programas de computador desenvolvidos para automatizar tarefas e simular interações entre humanos e sistemas computacionais.

De acordo com (Júnior e Carvalho, 2018), os *bots* são *softwares* criados para imitar ações humanas de forma repetitiva e realizar interações no contexto humano-computador. Segundo o autor, as primeiras versões desses sistemas apresentavam funcionamento mais simples e ofere-

ciam pouca capacidade de interação. Com a evolução tecnológica, esses programas passaram a incorporar mecanismos capazes de analisar previamente as necessidades dos usuários, permitindo uma interação mais natural.

Ainda conforme Júnior e Carvalho (2018), uma das principais características de um *bot* é a presença de um código desenvolvido para automatizar funções do cotidiano, especialmente aquelas relacionadas à interação com pessoas. Em alguns casos, esses sistemas podem inclusive executar atividades tipicamente humanas, simulando comportamentos durante a realização dessas tarefas. Nesse contexto, surgiram os *chatbots*, também conhecidos como *chatterbots*, que são robôs projetados especificamente para realizar conversações com usuários.

2.4 Geração Aumentada por Recuperação

A Inteligência Artificial Generativa, especialmente no contexto dos Grandes Modelos de Linguagem (LLMs) e da técnica de Recuperação Aumentada por Geração (RAG), tem sido apontada como um elemento de transformação nos modelos de Recuperação da Informação e nas ferramentas de busca.

De acordo com Gonçalves e Pavan (2023), a Inteligência Artificial Generativa, com foco em LLMs e RAG, representa uma mudança significativa nos modelos de recuperação da informação, ao integrar novas formas de geração e acesso ao conhecimento. Segundo os autores, a Ciência da Informação pode contribuir oferecendo bases conceituais que auxiliem no desenvolvimento de ambientes de busca mais efetivos. Na prova de conceito apresentada pelos autores, são discutidas e aproximadas as técnicas de RAG e LLM no contexto da Recuperação da Informação, demonstrando seu funcionamento e o potencial de aplicação dessas tecnologias para aprimorar tais ambientes.

No contexto técnico, o RAG utiliza tanto as informações armazenadas nos parâmetros do modelo durante o treinamento, conhecidas como memória paramétrica, quanto informações provenientes de fontes externas, chamadas de memória não paramétrica. A combinação dessas duas fontes de conhecimento permite que o sistema gere respostas de forma mais natural, utilizando o conhecimento previamente aprendido e dados recuperados dinamicamente para produzir respostas mais precisas e atualizadas.

Os Grandes Modelos de Linguagem (LLMs) revolucionaram a busca de informações, fazendo uma transição da exploração ativa para uma experiência mais passiva de recuperação de informação. No entanto, quando utilizados como fonte primária de conhecimento, os mo-

delos muitas vezes produzem conteúdo impreciso ou “alucinado” e carecem de atualizações de informações em tempo real, devido à dependência da memória paramétrica do modelo. O paradigma da Geração Aumentada por Recuperação (RAG) surgiu como uma solução para esses desafios de lidar com tarefas de intenso uso de conhecimento, combinando a robusta capacidade da recuperação de informação tradicional com a sofisticada geração de texto de LLMs. Este estudo propõe utilizar a arquitetura RAG para auxiliar na automatização da fase de levantamento bibliográfico para revisões de literatura, contando com uma recuperação de documentos relevantes aprimorada, com maior compreensão do conteúdo dos documentos, e uma experiência de usuário enriquecida por meio de sumarização do material e síntese da relevância dos resultados. O objetivo é simplificar o processo de busca, com resultados mais relevantes e permitindo uma pré-seleção mais eficiente, através das sumarizações (PROENÇA, 2024).

No contexto dos Grandes Modelos de Linguagem (LLMs), alucinação refere-se à geração de respostas que não se baseiam em informações factuais ou precisas. Elas podem ocorrer quando o modelo produz texto que inclui detalhes, fatos ou alegações que são fictícios, enganosos ou totalmente fabricados, em vez de fornecer informações confiáveis e verdadeiras. Em cenários em que é exigida extrema fidelidade e precisão nas respostas, como no direito, isso pode ser prejudicial, visto que qualquer erro pode causar uma grande confusão futuramente. Como os LLMs não possuem um entendimento semântico profundo, eles geram respostas com base em padrões linguísticos presentes nos dados utilizados para treinamento, o que pode levar à produção de informações incorretas ou enganosas (CASTRO, 2025).

O desenvolvimento do presente trabalho foi dividido em cinco fases metodológicas principais, que descrevem o processo desde a coleta dos dados brutos até a avaliação do sistema final. As fases foram: (1) Coleta e Preparação de Dados, (2) Criação da Base de Conhecimento, (3) Desenvolvimento do Sistema RAG, (4) Construção da Interface do Usuário e (5) Teste e Avaliação.

3.1 Coleta e Preparação de Dados

Os dados dos editais foram coletados de editais mais recentes e relevantes relacionados a pesquisas científicas, regulamento de graduação e assistência estudantil disponibilizados nos sites web do Instituto Federal Goiano - Campus Urutaí, priorizando arquivos em PDF. Os arquivos passaram por um processo de limpeza, onde foi removido cabeçalhos, figuras, rodapés e as tabelas foram simplificadas. Depois de testes com esses dados, foi necessário adicionar também resumos do conteúdo ao final dos PDF, o que resultou em respostas mais assertivas pelo *chatbot*.

3.2 Criação da Base de Conhecimento

Após a preparação, na segunda fase os dados foram transformados em um formato otimizado para recuperação. Os documentos pré-processados foram segmentados em *chunks* (segmentos menores e gerenciáveis de texto). Para cada *chunk*, foi gerado um vetor numérico (*embedding*) que representava seu significado semântico. Para essa finalidade, foi utilizado um modelo de *embedding* pré-treinado da plataforma Hugging Face, especificamente o modelo

"BAAI/bge-base-en-v1.5". Os *embeddings* e seus respectivos *chunks* de texto original foram armazenados em um banco de dados vetorial, onde foi criado um índice em memória utilizando a biblioteca LlamaIndex. Esta estrutura permitiu a busca eficiente por similaridade semântica durante a fase de recuperação.

3.3 Desenvolvimento do Sistema RAG

A terceira etapa envolveu a implementação do núcleo (core) do *chatbot*. A arquitetura desta fase foi inspirada na abordagem de Mangla (2024, tradução nossa). Dada uma consulta do usuário, o sistema realizava uma busca por similaridade de vetor no banco de dados vetorial para identificar os *chunks* de texto mais relevantes e contextualmente alinhados à pergunta. Os *chunks* recuperados foram concatenados com a pergunta original do usuário e enviados como *prompt* para o Grande Modelo de Linguagem (LLM), especificamente o modelo TinyLlama/TinyLlama-1.1B-Chat-v1.0. O LLM foi configurado para gerar respostas concisas e informativas com base exclusivamente no contexto fornecido pelos *chunks* recuperados, minimizando o risco de "alucinações".

3.4 Construção da Interface do Usuário

Na quarta fase do projeto, foi desenvolvida a interface gráfica do usuário para permitir a interação com o sistema RAG (descrito na Fase 3.3). Para esta finalidade, foi utilizada a biblioteca Gradio. Segundo a página oficial, o Gradio é um pacote Python de código aberto que permite criar rapidamente uma demonstração ou aplicação web para seu modelo de aprendizado de máquina, API ou qualquer função Python, sem necessidade de experiência com JavaScript, CSS ou hospedagem web (GRADIO, 2025, tradução nossa). Com base nisso, foi implementada uma interface de chat, composta por um campo de entrada de texto (para a pergunta do usuário) e um campo de saída (para a resposta gerada pelo chatbot).

3.5 Teste e Avaliação

A quinta fase da metodologia consistiu em um ciclo de teste e avaliação qualitativo e iterativo, focado primariamente na validação da fonte de dados. Para garantir a qualidade do banco de conhecimento antes de sua unificação, foi adotada uma abordagem de testes individuais. Para cada PDF editado, foi criado um índice vetorial temporário, e o *chatbot* foi testado com

perguntas direcionadas àquele documento específico. Inicialmente, observou-se que as respostas do *chatbot*, baseadas apenas nos PDFs com limpeza de formatação, eram frequentemente insatisfatórias. Para solucionar esta deficiência, o processo de preparação de dados refinados: foram adicionados pequenos resumos do conteúdo ao final de cada arquivo PDF. Este ciclo de "testar-e-refinar" foi repetido para cada documento, e a inclusão dos resumos demonstrou uma melhora significativa na assertividade das respostas. Somente após essa validação individual, todos os arquivos PDF refinados foram unificados para a criação do banco de dados vetorial final.

3.6 Ferramentas e recursos utilizados

Todo o desenvolvimento do projeto, desde a preparação dos dados até a implementação do *chatbot* e da interface, foi realizado utilizando o ambiente Google Colabory. O Google Colabory, de acordo com a própria documentação, é uma ferramenta que permite escrever e executar Python no navegador sem nenhuma configuração necessária. Além disso, possui acesso a GPUs e compartilhamento fácil (GOOGLE, 2025, tradução nossa). Os notebooks do Colab permitem combinar código executável e *rich text* em um só documento, além de imagens, HTML, LaTeX e muito mais. No contexto deste trabalho, os notebooks foram armazenados na conta Google Drive do pesquisador, facilitando o versionamento e o compartilhamento. Foi possível importar os dados (editais em PDF) diretamente do Google Drive, bem como do GitHub.

Outra ferramenta utilizada foi o LlamaIndex, que é uma ferramenta poderosa projetada para aprimorar ainda mais as capacidades de aplicativos baseados em RAG (Recuperação, Atribuição e Geração). O LlamaIndex oferece uma maneira integrada de combinar modelos de recuperação e geração para construir aplicativos baseados em RAG. Foi utilizado para a orquestração de todo o fluxo RAG, criação do índice vetorial e conexão com os modelos.

A plataforma Hugging Face foi utilizada como o principal repositório para os modelos de inteligência artificial. Por meio de suas bibliotecas (transformers, sentence-transformers), foi feito o carregamento dos dois modelos centrais do projeto: o modelo de *embedding* BAAI/bge-base-en-v1.5 (utilizado na Fase 3.2) e o Grande Modelo de Linguagem (LLM) TinyLlama-1.1B-Chat-v1.0 (utilizado na Fase 3.3). Por fim, para a construção da interface de usuário (Fase 3.4), foi utilizada a biblioteca Gradio.

CAPÍTULO 4

RESULTADOS E DISCUSSÃO

4.1 Apresentação da Interface do Usuário

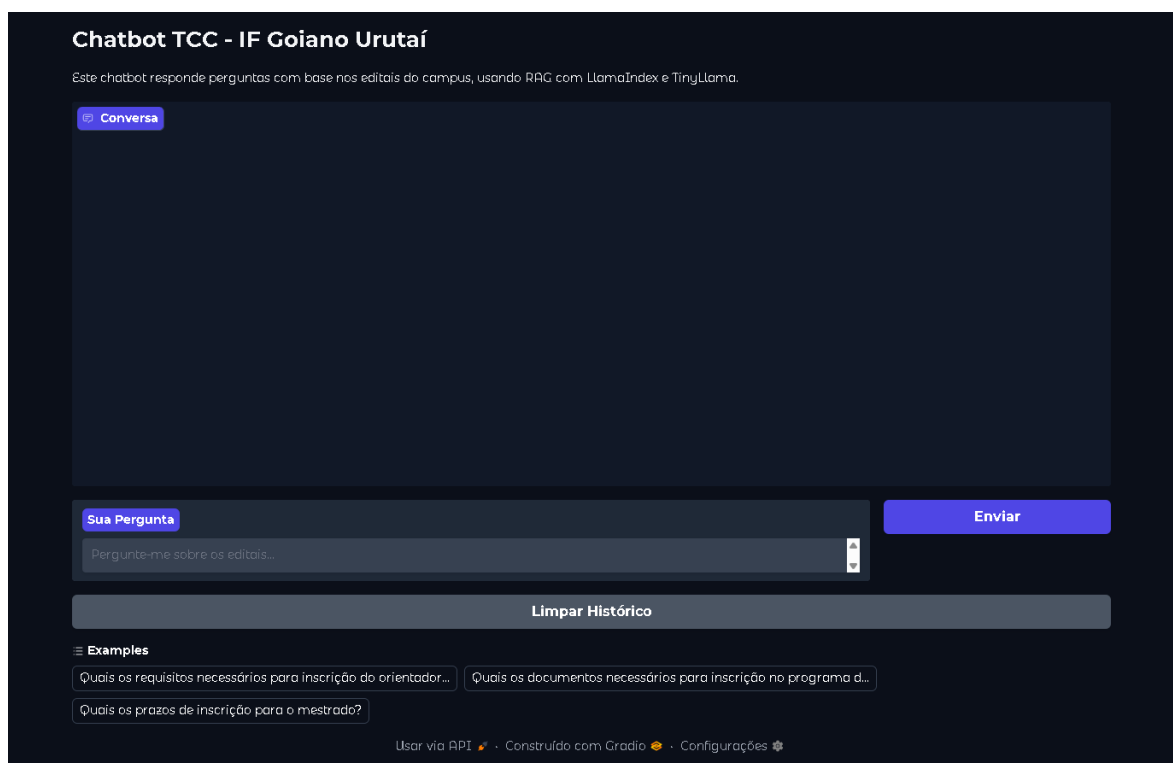


Figura 4.1: Interface de Usuário do Chatbot.

A interface, Figura 4.1, implementada com a biblioteca Gradio, apresenta um design limpo e funcional. No topo, exibe o título do projeto ("Chatbot TCC - IF Goiano Urutaí") e uma breve descrição de sua funcionalidade ("Este *chatbot* responde perguntas com base nos editais do campus usando RAG com LhmaIndex e TinyLhama"). A área principal é composta por uma janela de diálogo para exibir a conversa, um campo de entrada de texto inferior ("Sua

Pergunta"), um botão de envio e um botão para "Limpar Histórico". Notavelmente, a interface também inclui uma seção de "Exemplos" ("Quais os requisitos necessários para inscrição do orientador", "Quais os documentos necessários para inscrição no programa"), que serve para guiar o usuário sobre o tipo de pergunta que o sistema é capaz de responder, facilitando o primeiro uso.

4.2 Análise de Resultados

Para validar o funcionamento prático do sistema, foram realizados testes com perguntas diretas baseadas no conteúdo dos editais. A Figura demonstra um caso de sucesso, onde foi submetida ao *chatbot* a pergunta: "O que deve conter o projeto de pesquisa?".

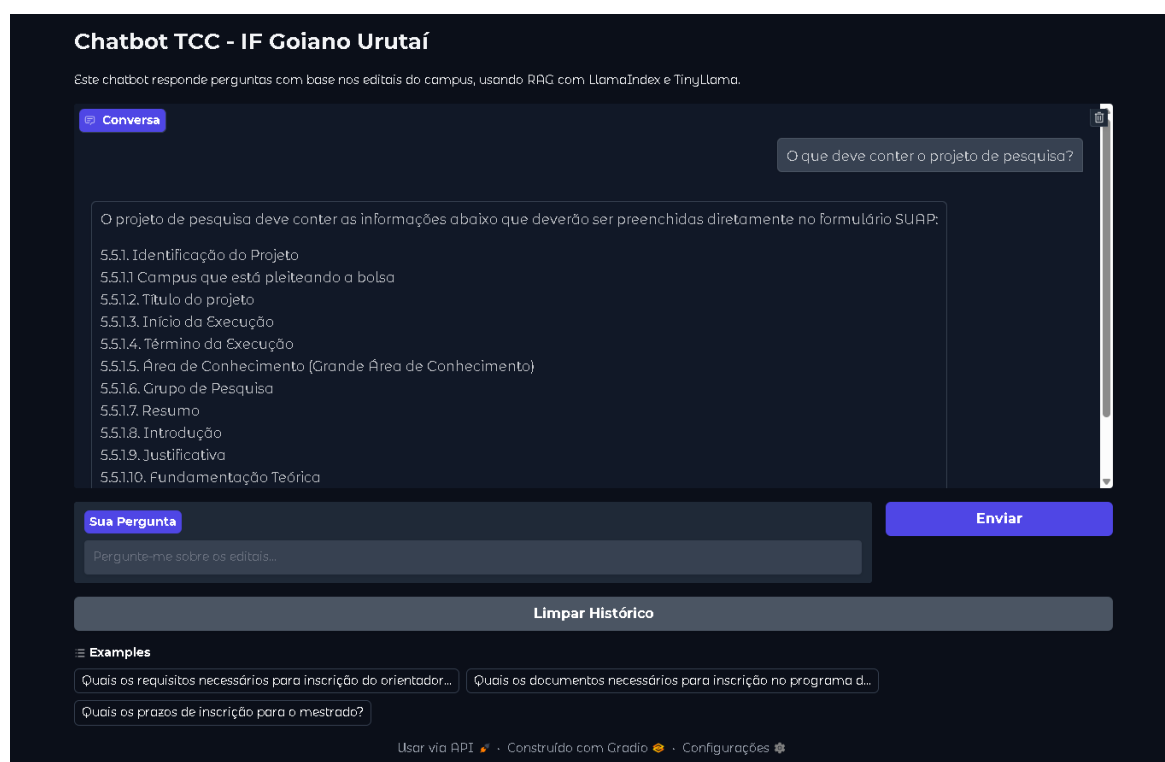


Figura 4.2: Resposta Correta em relação à Projeto de Pesquisa.

Como demonstrado na Figura 4.2, o *chatbot* respondeu com uma lista estruturada e detalhada, extraída diretamente do edital correspondente. A resposta cita os itens exatos que devem constar no formulário SUAP, como "5.5.1 Identificação do Projeto", "5.5.1.2 Título do projeto", "5.5.1.7 Resumo", "5.5.1.11 Objetivo Geral", entre outros. Este resultado é um caso de sucesso significativo por dois motivos principais, pois mostrou que o sistema RAG foi capaz de buscar no banco de dados vetorial e localizar o *chunk* (pedaço de texto) exato que continha a resposta para a pergunta do usuário e que o LLM cumpriu sua instrução de gerar uma resposta

baseada exclusivamente no contexto recuperado, reproduzindo a lista de forma precisa, sem inventar informações. Isso demonstra a metodologia empregada, desde a preparação dos dados até a arquitetura RAG.

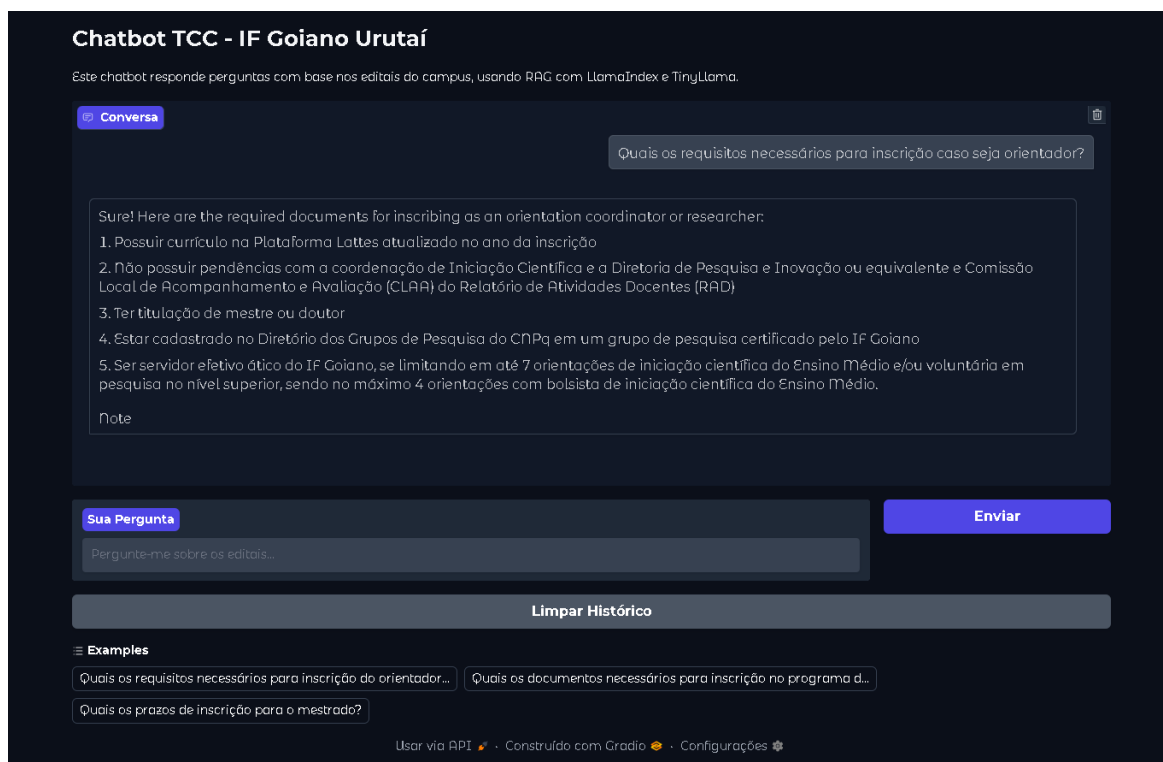


Figura 4.3: Resposta Correta em relação aos Requisitos do Orientador.

Esse segundo exemplo na figura 4.3 também é um resultado positivo da aplicação, demonstrando como o sistema RAG foi capaz de buscar no banco de dados vetorial e localizar o *chunk* (pedaço de texto) exato que continha a resposta para a pergunta do usuário e que o LLM cumpriu sua instrução de gerar uma resposta baseada exclusivamente no contexto recuperado, reproduzindo a lista de forma precisa, sem inventar informações.

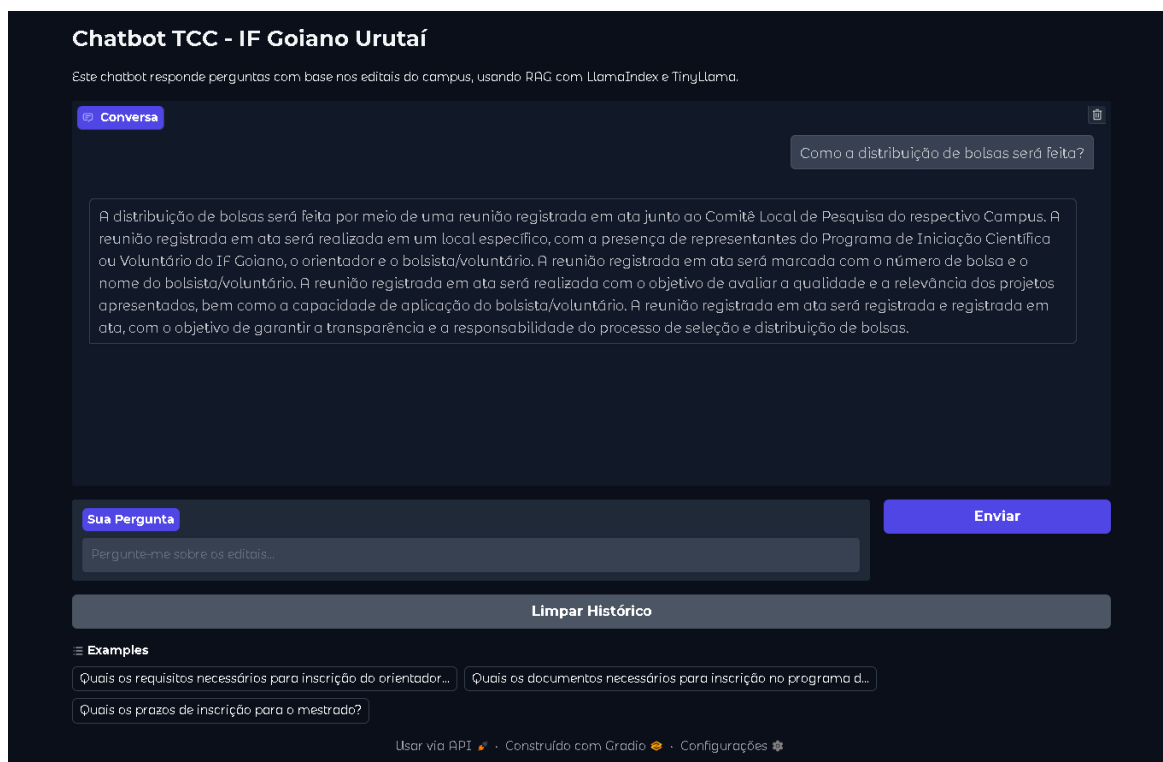


Figura 4.4: Resposta Correta relacionado à Distribuição de Bolsas.

Terceiro exemplo de resultado positivo na 4.4, agora em relação ao funcionamento da distribuição de bolsas. A pergunta simples: 'Como a distribuição de bolsas será feita?' obteve uma resposta coesa que se baseou corretamente na busca no banco de dados vetorial.

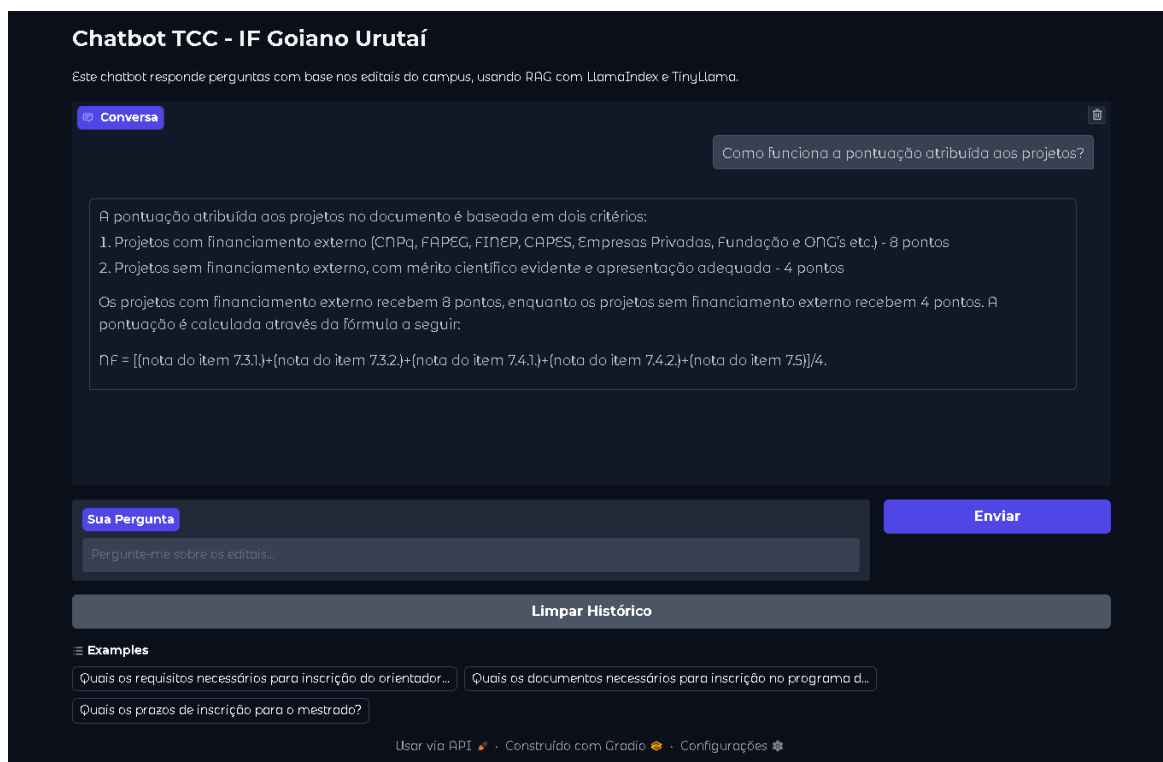


Figura 4.5: Resultado Positivo relacionado à Pontuação de Projetos.

Quarto exemplo de resultado positivo na 4.5, onde o questionamento foi: "Como funciona a pontuação atribuída aos projetos?". O *chatbot* trouxe resultado detalhado da divisão de pontos, cálculo utilizado e demonstração da fórmula.

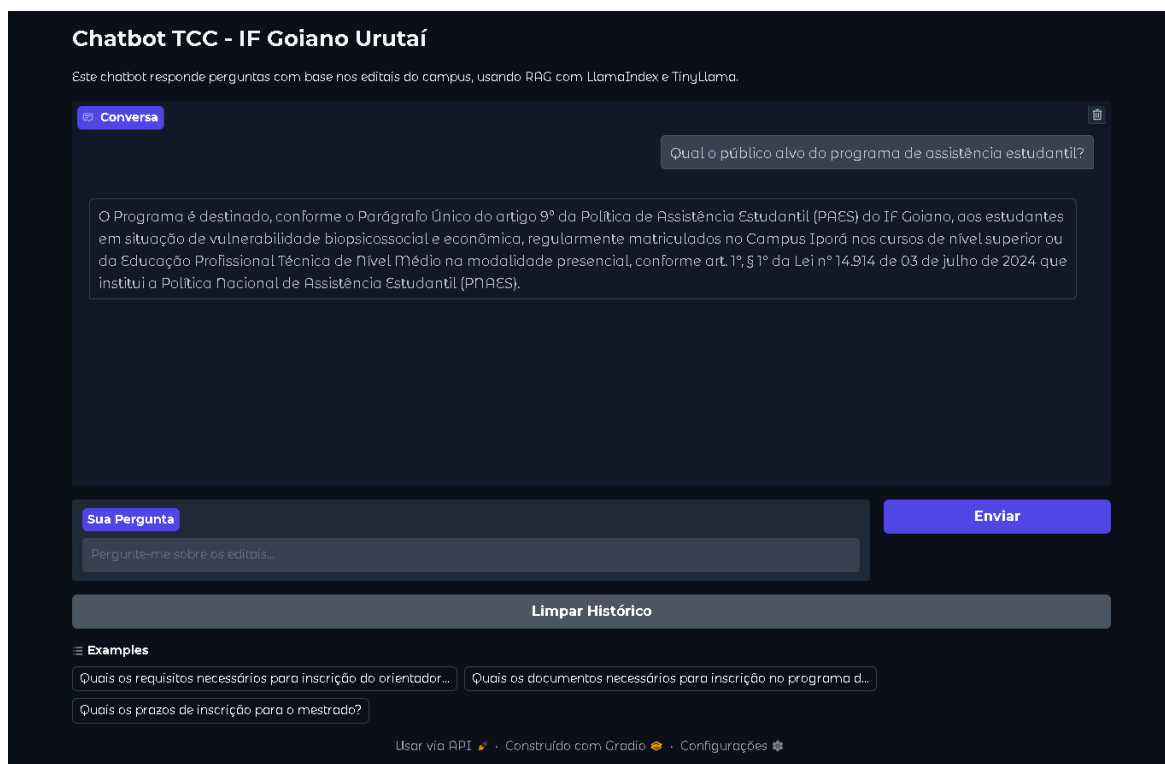


Figura 4.6: Resultado Positivo relacionado ao Programa de Assistência Estudantil.

Quinto exemplo de resultado positivo na **4.6**, onde o questionamento foi: "Qual o público alvo do programa de assistência estudantil?". O *chatbot* trouxe resultado correto com base nos editais relacionados a assistência estudantil presentes em seu banco de dados.

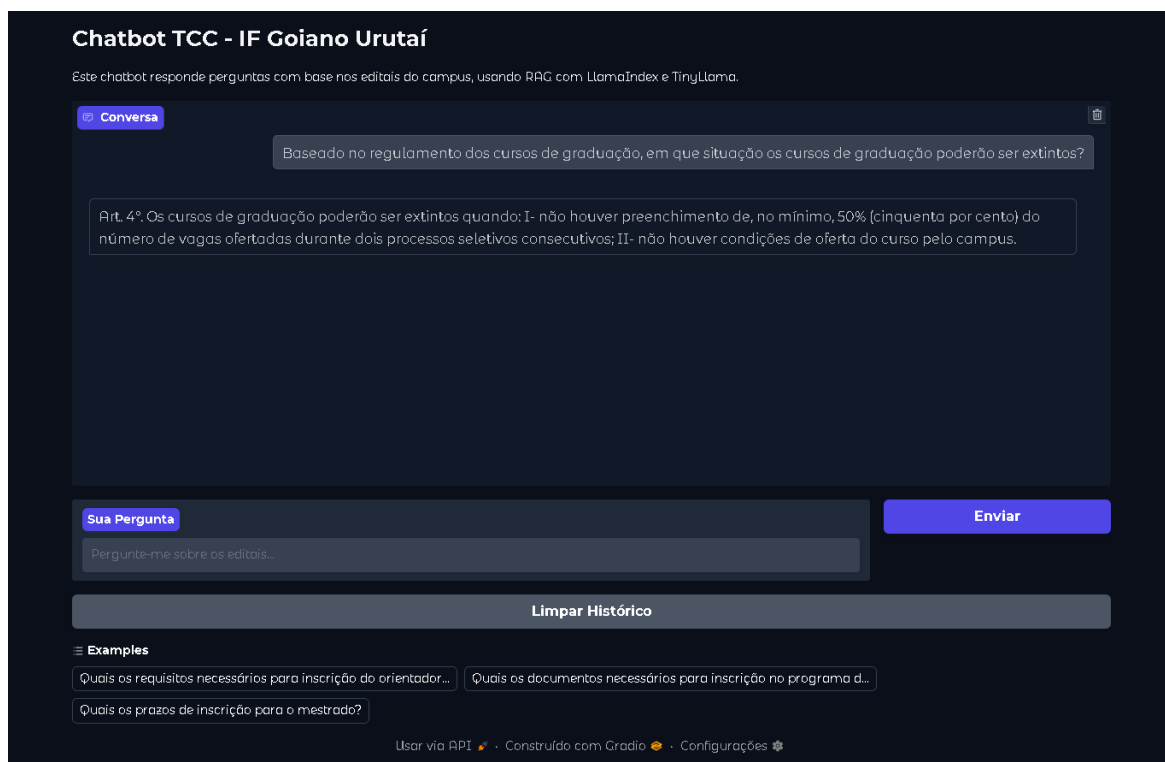


Figura 4.7: Resultado Positivo relacionado ao Regulamento de Cursos de Graduação.

Sexto exemplo de resultado positivo na 4.7, onde o questionamento foi: "Baseado no regulamento dos cursos de graduação, em que situação os cursos de graduação poderão ser extintos". O *chatbot* respondeu corretamente citando as condições nas quais os cursos poderão ser extintos.

4.3 Discussão: Limitações e o Processo de Refinamento

Uma parte crucial da discussão dos resultados não é apenas analisar o produto final, mas também o processo de desenvolvimento. Conforme descrito na Fase 3.5 da Metodologia, os testes iniciais revelaram limitações significativas que exigiram o refinamento dos dados. A Figura 4.8 é um exemplo de um desses testes, realizado diretamente no notebook de desenvolvimento (Google Colab), antes da implementação da interface Gradio.

```

)

# 2. Definição do Prompt
prompt = "A inscrição para participar da avaliação A Comissão de Heteroidentificação do IF Goiano ocorrerá que dia?" # A pergunta ou início de texto que o modelo deve

# 3. Execução da Geração de Texto
sequences = pipe(
    prompt, # O prompt de entrada
    max_new_tokens=100, # Número máximo de novos tokens (palavras/subpalavras) a serem gerados
    do_sample=True, # Ativa o amostragem estocástica (saída mais criativa/variação)
    top_k=10, # Limita a amostragem aos 10 tokens mais prováveis em cada etapa
    return_full_text = False, # Retorna apenas o texto gerado, sem o prompt original incluído
)

# 4. Exibição do Resultado
for seq in sequences: # Itera sobre as sequências geradas (geralmente uma, a menos que n_sequences seja maior que 1)
    print(f"Result: {seq['generated_text']}") # Imprime o texto gerado

specified. 'device' will override 'device_map'. You will most likely encounter unexpected behavior. Please remove 'device' and keep 'device_map'.

io A da inscrição para participar da avaliação A de Goiás da Comissão de Heteroidentificação do IF Goiano, diz que a avaliação será realizada no dia 8 de junho de 2022.

```

Figura 4.8: Exemplo de resposta incorreta.

Na Figura 4.8, o *prompt* enviado ao sistema foi: "A inscrição para participar da avaliação A Comissão de Heteroidentificação do IF Goiano ocorrerá que dia?". O modelo, em vez de fornecer uma resposta clara, retornou um fragmento de texto incompleto, mal formatado e cita uma data incorreta. Este tipo de resposta insatisfatória era comum nas fases iniciais desta pesquisa.

```

[17] 3min while True:
        query = input("<|usuário|>\n")
        if query == 'parar':
            break
        response = query_engine.chat(query)
        print("\n", response, "\n")

... Seu chat está pronto para conversar, caso queira encerrar a conversa, digite 'parar'
<|usuário|>
Quais os requisitos necessários para inscrição do orientador no programa de bolsa de iniciação científica?

Sure! Here are the required documents for inscribing as an orientador in the bursary program of the Institute of Federal Goiania (IF Goiano):

1. A valid academic degree (Master's or PhD) from an accredited university or research institution.
2. A minimum of 5 years of relevant experience in the field of research or academic work.
3. A record of publications in high-impact scientific journals or conferences.
4. A record of presenting research at national and international conferences.
5. A record of teaching experience in higher education

<|usuário|>
em português

Sure, here's the translated version of the instruction:

Para inscrição no programa de bolsa de iniciação científica, é necessário apresentar os seguintes documentos:

1. Mestrado ou doutorado atualizado no ano da inscrição;
2. Não possuir pendências com a coordenação de iniciação científica e a diretoria de pesquisa e inovação ou equivalente e Comissão Local de Acompanhamento e A

<|usuário|>
parar

```

Figura 4.9: Segundo exemplo de resposta incorreta.

Na figura 4.9, o *prompt* dado ao sistema foi: "Quais os requisitos necessários para a inscrição do orientador no programa de bolsa de iniciação científica?". O modelo retornou uma resposta em inglês, mesmo que a pergunta tenha sido feita em português. Além disso, a resposta estava incorreta e foi necessário pedir a mudança para o português para obter a resposta no idioma correto. A análise desses resultados (conforme descrito na Metodologia, Fase 3.5) indicou que o sistema RAG estava recuperando *chunks* de texto corretos, mas que careciam de contexto suficiente para o LLM (TinyLlama) formular uma resposta coesa. Esta

constatação validou a decisão metodológica de refinar os dados. A solução, descrita na Fase 3.1 e 3.5, foi a de adicionar resumos contextuais ao final de cada arquivo PDF. Portanto, a "falha" demonstrada na Figura 4.8 é apresentada aqui para comprovar o sucesso do processo de refinamento iterativo. O resultado positivo visto na Figura 4.2 (o teste no Gradio) é uma consequência direta da melhoria nos dados, que mitigou drasticamente a ocorrência de respostas incompletas ou fragmentadas na versão final do *chatbot*.

Este Trabalho de Conclusão de Curso se propôs a desenvolver e avaliar um *chatbot* baseado na arquitetura RAG (Geração Aumentada por Recuperação), com o objetivo de otimizar o acesso às informações contidas nos editais do IF Goiano Campus Urutaí. Diante do volume e da complexidade desses documentos a pesquisa visa reduzir o tempo de consulta de alunos e professores e diminuir a carga administrativa das secretarias. Conclui-se que o objetivo geral foi plenamente alcançado. O protótipo desenvolvido, utilizando LlamaIndex para orquestração, TinyLlama como LLM e Gradio para interface, demonstrou ser uma solução viável e funcional. Conforme apresentado no Capítulo 4, o *chatbot* foi capaz de responder com precisão a perguntas diretas sobre os editais.

A principal contribuição deste trabalho, no entanto, não reside apenas na implementação da ferramenta, mas na validação da metodologia de preparação de dados. Os testes iniciais revelaram que a simples extração de texto de arquivos PDF é insuficiente para um sistema RAG de qualidade. A etapa de refinamento, que envolveu a adição de resumos contextuais ao final de cada documento, mostrou-se uma ação metodológica crucial e foi a principal responsável pelo sucesso do protótipo final. Apesar dos resultados positivos, o estudo possui limitações. O sistema depende de um LLM pequeno (TinyLlama) para operar sem custos de API, o que limita sua capacidade de inferência em perguntas mais complexas. Além disso, o processo de preparação de dados (limpeza e adição de resumos) é manual e precisa ser refeito a cada novo edital publicado.

REFERÊNCIAS BIBLIOGRÁFICAS

CARVALHO, M. C. P. d. *Recuperação de Informação e Processamento de Linguagem Natural: Mini-curso*. [S.l.], 2003. Mini-curso, PDF Disponível em. Disponível em: <https://d1wqtxts1xzle7.cloudfront.net/43022678/minicurso-jaia2003-libre.pdf>.

CASELI, H. d. M.; NUNES, M. d. G. V.; PAGANO, A. S. O que é pln? *Processamento de linguagem natural: conceitos, técnicas e aplicações em português*, 2023.

CASTRO, R. C. F. d. *O uso da geração aumentada por recuperação na validação da extração de argumentos de documentos jurídicos*. Dissertação (Trabalho de Conclusão de Curso) — Universidade Federal do Ceará, 2025. Orientador: Magalhães, Regis Pires. Disponível em: <http://repositorio.ufc.br/handle/riufc/82854>.

GONÇALVES, L. H.; PAVAN, T. R. *Recuperação de Informação com Inteligência Artificial Generativa: um Estudo Sobre a Geração Aumentada por Recuperação (RAG)*. 2023. Disponível em: <https://riut.utfpr.edu.br/jspui/bitstream/1/36353/1/recuperacaoinformacaointelligenceartificialgenerativa.pdf>.

GOOGLE. *Google Colaboratory*. 2025. <https://colab.research.google.com/>. Acessado em 12 de novembro de 2025.

GRADIO. *Quickstart - Gradio*. 2025. <https://www.gradio.app/guides/quickstart>. Acessado em 11 de novembro de 2025.

JÚNIOR, C. F. d. C.; CARVALHO, K. R. S. d. A. d. Chatbot: uma visão geral sobre aplicações inteligentes. *Revista Sítio Novo*, v. 2, n. 2, p. 68–84, 2018. Disponível em: <https://sitionovo.ifo.edu.br/index.php/sitionovo/article/view/140>.

MANGLA, P. Llamaindex: Building a smarter rag-based chatbot. In: CHUGH, P.; GOSTHIPATY, A. R.; HUOT, S.; KIDRIAVSTEVA, K.; RAHA, R. (Ed.). *PyImageSearch*. [s.n.], 2024. Disponível em: <https://pyimg.co/0jvku>.

PROENÇA, V. L. *Automatização da revisão de literatura científica com geração aumentada por recuperação*. Dissertação (Trabalho de Conclusão de Curso de Graduação em Ciência da Computação) — Universidade Federal do Rio de Janeiro, 2024. Orientador: Silva, João Carlos Pereira da. Disponível em: <https://pantheon.ufrj.br/handle/11422/23629>.

SOUSA, A. L. C. d.; SANTOS, F. M. d.; RODRIGUES, A. L. S. *A Inteligência Artificial e a Ciência da Informação: uma análise do uso do ChatGPT como ferramenta de Recuperação da*

Informação. 2023. Disponível em: <<https://periodicos.educacaotransversal.com.br/index.php/rechso/article/view/90/94>>.

VARAS, J.; CORONEL, B. V.; VILLAGRÁN, I.; ESCALONA, G.; HERNANDEZ, R.; SCHUIT, G.; DURÁN, V.; LAGOS-VILLASECA, A.; JARRY, C.; NEYEM, A.; ACHURRA, P. Inovações no treinamento cirúrgico: explorando o papel da inteligência artificial e dos grandes modelos de linguagem (LLM). *Revista do Colégio Brasileiro de Cirurgias*, v. 50, p. e20233605, 2023. Disponível em: <<https://www.scielo.br/j/rcbc/a/KVsSNSfHQhkGJjDMhkhwrKk/?lang=pt>>.

XIAO TONG ZHU, J. *Foundations of Large Language Models*. 2025. Disponível em: <<https://arxiv.org/pdf/2501.09223>>.