Um LLM com RAG para as Normativas Institucionais da UDESC

Lucas Pietro Biasi Rayzer[1], **Fernando Santos**[1]

[1]Departamento de Engenharia de Software
Universidade do Estado de Santa Catarina (UDESC)
Ibirama – SC – Brazil

lucas.rayzer7733@edu.udesc.br, fernando.santos@udesc.br

Abstract. *The regulations of the Universidade do Estado de Santa Catarina (UDESC) are made available to students and staff in a dispersed manner. This hinders the quick retrieval of regulations that establish rules for everyday situations, such as taking makeup exams and requesting financial resources. This article describes an ongoing project aimed at providing a Large Language Model (LLM) combined with Retrieval-Augmented Generation (RAG) to query institutional regulations through a chatbot. The goal is to optimize access to regulatory information, thereby transforming the experience of students, faculty, and staff.*

Resumo. *As normativas da Universidade do Estado de Santa Catarina (UDESC) são disponibilizadas aos estudantes e servidores de forma dispersa. Isto dificulta a rápida localização de normativas que estabelecem regras para situações quotidianas, tais como realização de avaliações em segunda chamada e solicitação de recursos financeiros. Este artigo descreve um trabalho em andamento cujo objetivo é disponibilizar um Modelo de Linguagem de Larga Escala (LLM) combinado com* Retrieval-Augumented Generation *(RAG) para consultar as normativas institucionais através de um* chatbot*. Com isso, pretende-se otimizar o acesso às informações normativas, transformando a experiência de estudantes, professores e servidores.*

1. Introdução

Atualmente, diversas instituições públicas buscam soluções tecnológicas capazes de otimizar processos. Exemplos incluem *chatbots* para esclarecimento de dúvidas e suporte a alunos e servidores [Oliveira 2024], e para agilizar o acesso a documentos públicos municipais [Rede de Acesso à Informação Aberta 2025]. Estes exemplos evidenciam o potencial dos agentes virtuais. No contexto universitário, a gestão de documentos internos é um desafio recorrente, já que o volume e a dispersão das informações comprometem a eficiência administrativa, dificultando o acesso a regulamentos institucionais.

Na Universidade do Estado de Santa Catarina (UDESC) a dispersão das normativas afeta discentes, docentes e servidores: estudantes enfrentam dificuldades para localizar informações essenciais e setores administrativos sofrem sobrecarga de atendimentos repetitivos. Esse cenário também impacta políticas de permanência estudantil, cujas informações, apesar de divulgadas, muitas vezes não chegam de forma clara ao público-alvo. As normativas da UDESC são estabelecidas pelos conselhos e câmaras da universidade. Anualmente, dezenas de normas são criadas ou atualizadas. Por exemplo, apenas o Conselho Universitário (CONSUNI - órgão superior da universidade) definiu 214

resoluções nos anos de 2023 a 2025, evidenciando o grande volume de normativas que determinam o funcionamento da universidade e que podem não ser conhecidas por seu público. Além dos conselhos e câmaras universitários, os diversos centros de ensino da UDESC podem emitir normativas próprias. Isso acaba tornando as normativas ainda mais específicas e de difícil acesso ao público, que pode enfrentar dificuldades para localizar rapidamente normativas envolvidas com situações quotidianas, tais como avaliações em segunda chamada e solicitação de recursos para participação em eventos.

Ao analisar os trabalhos relacionados, a aplicação de modelos de linguagem combinados com a abordagem RAG para otimizar o acesso a informações em contextos específicos foi um ponto em comum entre os estudos analisados. O estudo de [Oliveira 2024] mostra como a integração entre raspagem de dados, LangChain e Gemini 1.5 Pro pode estruturar um *chatbot* voltado ao acesso acadêmico. Já [Saha and Saha 2024] expandem essa perspectiva para o suporte de estudantes internacionais, utilizando dados de comunidades virtuais e métricas robustas para avaliação. Esses trabalhos reforçam a relevância dos LLMs aliados ao RAG em cenários de apoio informacional, mas nenhum deles foca na centralização de normativas institucionais dispersas.

Este artigo descreve um trabalho em andamento cujo objetivo é o desenvolvimento de um *chatbot* baseado em inteligência artificial para centralizar e disponibilizar informações institucionais da UDESC. Com isto pretende-se aumentar a autonomia acadêmica, reduzindo a demanda por atendimentos presenciais. O trabalho prevê a coleta e processamento de documentos oficiais dos conselhos universitários, extração textual, normalização e armazenamento em banco vetorial, possibilitando o uso de técnicas de recuperação de informação aliadas a um LLM. O objetivo central é oferecer respostas rápidas e precisas sobre processos burocráticos, enquanto objetivos específicos incluem mapear documentos institucionais, treinar o modelo com dados reais e avaliar o desempenho da solução em termos de usabilidade, precisão e clareza. Essa abordagem justifica-se pela necessidade crescente de transparência e acessibilidade nos trâmites internos, acompanhando o movimento global de digitalização acadêmica e administrativa.

A proposta se apoia na fundamentação teórica sobre *Natural Language Processing* (NLP) e *Retrieval-Augumented Generation* (RAG). O NLP permite que máquinas compreendam e gerem linguagem humana, dividindo-se em etapas como análise sintática e semântica para interpretar o que o usuário deseja e produzir respostas adequadas [Martins et al. 2020, Caseli and Nunes 2024]. A abordagem RAG complementa LLM's, cujo conhecimento é estático e limitado aos dados de seu treinamento, ao permitir que eles acessem e incorporem informações externas e atualizadas de uma base de documentos em tempo real para gerar respostas contextualizadas [Arslan et al. 2024]. Com base em trabalhos correlatos que demonstram a eficácia da arquitetura RAG em cenários similares [Oliveira 2024], projeta-se que a combinação dessas tecnologias resultará em um sistema robusto, transparente e eficiente, capaz de otimizar o acesso às informações acadêmico-administrativas e transformar a experiência de estudantes, professores e servidores.

2. Desenvolvimento do Trabalho

O desenvolvimento do trabalho envolve etapas de seleção de LLM adequado, combinação deste LLM com RAG, e disponibilização e validação do *chatbot* junto aos usuários.

A seleção do LLM será realizada a partir de modelos disponíveis no repositório

Hugging Face[1]. Serão considerados aspectos relacionados aos parâmetros dos LLMs, capacidade de processamento de *tokens* e licenças de uso, a fim de identificar qual deles apresenta melhor equilíbrio entre desempenho, consumo de recursos computacionais e qualidade de respostas. A Tabela 1 apresenta os LLMs pre-selecionados no Hugging Face cujas características serão analisadas para verificar a aderência aos critérios relacionados. Entre eles se destacam o LLaMA 3 por boa performance, tendo modelos com diferentes números parâmetros (8 ou 70 bilhões). Os modelos Mixtral e Gemma, que possuem menor quantidade de parâmetros (7 bilhões), foram adicionados a lista para avaliação. A avaliação será realizada em máquina virtual disponibilizado pelo setor de tecnologia da UDESC, composta por processador Intel Xeon Platinum, 64 GB de memória RAM e GPU Nvidia L40S-48Q, com 43 GB vRam e 568 *Tensor Cores* (4ª geração, Ada Lovelace)[2].

Tabela 1. LLMs pré-selecionados para o trabalho

Modelo	Nome no Hugging Face	Licença
LLaMA 3 - 70B Instruct	`meta-llama/Llama-3.3-70B-Instruct`	Llama 3.3
LLaMA 3 - 8B	`meta-llama/Llama-3.1-8B-Instruct`	Llama 3.1
Mixtral 8x7B Instruct	`mistralai/Mixtral-8x7B-Instruct-v0.1`	Apache 2.0
Gemma 7B Instruction	`google/gemma-7b-it`	Gemma RL

Paralelamente, serão coletados documentos institucionais oficiais da UDESC, como atas de conselhos, resoluções, regulamentos e editais. Os arquivos em formato PDF serão processados com uma biblioteca de leitura e carregamento de PDF's, como o PDFPlumber[3], permitindo a extração de texto e a organização dos dados de acordo com metadados relevantes, como data, órgão emissor, tipo de documento e assunto. Após a extração, os textos passarão por uma etapa de pré-processamento, que inclui a remoção de caracteres especiais, rasuras, espaços em excesso e quebras de linha indevidas, garantindo a consistência do *corpus*.

Concluída a normalização textual, os dados serão submetidos à etapa de vetorização por meio da geração de *embeddings*, possibilitando uma representação em formato numérico. Para isso, serão utilizadas bibliotecas como *sentence-transformers*. Os vetores resultantes serão armazenados em um banco de dados vetorial, de modo a permitir buscas rápidas e relevantes. Antes da vetorização, os documentos serão submetidos a um processo de *chunking* em que cada texto é dividido em partes menores e semanticamente coerentes. Essa segmentação tem como objetivo otimizar a recuperação de informações pelo mecanismo de busca, melhorar a qualidade das representações vetoriais e reduzir a perda de contexto durante o processo de inferência do modelo de linguagem.

Em seguida, será implementado o sistema RAG nos LLMs pre-selecionados. Os documentos vetorizados servirão de base para a recuperação de trechos relevantes, que serão enviados ao modelo de linguagem para a geração de respostas personalizadas. Essa integração será realizada com auxílio de *frameworks* como LangChain[4] ou Haystack[5], que permitem orquestrar a comunicação entre o banco vetorial e o modelo de linguagem.

[1]https://huggingface.co/

[2]https://www.nvidia.com/pt-br/data-center/l40s/

[3]https://python.langchain.com/docs/integrations/document_loaders/pdfplumber/

[4]https://www.langchain.com/

[5]https://haystack.deepset.ai/

Para a validação do sistema, será realizado um experimento considerando estudantes e servidores da universidade. No experimento, os participantes receberão formulários contendo perguntas sobre trâmites universitários. Os participantes interagirão com o *chatbot*, e as respostas fornecidas serão avaliadas de acordo com critérios de relevância, precisão e clareza. A partir dessa análise, será possível avaliar os modelos ao que diz respeito à qualidade das respostas, velocidade de processamento e consumo de recursos.

3. Considerações Finais

O trabalho proposto tem como objetivo central facilitar o acesso a informações normativas da UDESC, promovendo autonomia para estudantes, professores e servidores no trato com trâmites acadêmicos e administrativos. Diferentemente de iniciativas correlatas, que priorizam conteúdos culturais ou informações gerais, esta proposta foca na unificação de documentos normativos estabelecidos pelos conselhos universitários, fornecendo respostas contextualizadas e confiáveis ao público universitário.

O trabalho encontra-se em desenvolvimento. Até o momento já foram coletados os documentos institucionais junto a universidade. Além disso, já foram analisadas ferramentas que poderão auxiliar no processo de integração e tratamento dos dados, bem como os LLMs que serão avaliados. As próximas etapas do desenvolvimento são a extração e tratamento dos textos coletados, seguido da codificação e integração com a abordagem RAG para viabilizar o funcionamento do *chatbot*. agente.

O diferencial do projeto está no processamento cuidadoso dos documentos oficiais, em especial arquivos em PDF, que exigem etapas de extração, normalização e vetorização antes de integrarem o sistema baseado em RAG. Espera-se que essa atenção no tratamento dos dados contribua para a precisão e clareza das respostas do *chatbot*.

Fernando Santos agradece à FAPESC pelo apoio financeiro recebido (TO2023TR246).

Referências

Arslan, M., Ghanem, H., Munawar, S., and Cruz, C. (2024). A survey on RAG with LLMs. *Procedia computer science*, 246:3781–3790.

Caseli, H. d. M. and Nunes, M. d. G. V. (2024). *Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português*. BPLN, São Carlos, 2 edition.

Martins, C. O., Lenz, M. L., Silva, M. B. F. D., et al. (2020). *Processamentos de Linguagem Natural*. SAGAH, Porto Alegre.

Oliveira, L. F. (2024). *Chatbot como ferramenta de apoio ao acesso às informações acadêmicas da UFSM*. Trabalho de conclusão de curso, Universidade Federal de Santa Maria (UFSM).

Rede de Acesso à Informação Aberta (2025). Chat diário oficial. Disponível em: `https://grupo-raia.org/projects/project/1`. Acesso em: 03 jun. 2025.

Saha, B. and Saha, U. (2024). Enhancing international graduate student experience through ai-driven support systems: A llm and rag-based approach. In *2024 International Conference on Data Science and Its Applications (ICoDSA)*, pages 300–304. IEEE.