EXPLORANDO O USO DE LLMS COM RAG PARA SIMPLIFICAR O ACESSO ÀS NORMATIVAS DA UNIVERSIDADE DO ESTADO DE SANTA CATARINA

https://doi.org/10.5902/2318133895063

Lucas Pietro Biasi Rayzer [1]
Fernando Santos[2]

Resumo
A dispersão de normativas acadêmicas dificulta o acesso a informações em universidades. Este estudo tem como objetivo investigar o uso de grandes modelos de linguagem – LLM – integrados à técnica retrieval-augmented generation – RAG – para simplificar o acesso às normativas da Universidade do Estado de Santa Catarina. Informações recuperadas a partir de documentos normativos foram utilizadas como contexto RAG em perguntas enviadas a dois LLMs: Meta Llama 3.1 e Google Gemini 2.5. A avaliação, baseada no framework Ragas, indicou que a qualidade do contexto recuperado é essencial para respostas fiéis e corretas. O Gemini apresentou maior robustez frente a contextos ruidosos, alcançando melhores índices de fidelidade e correção, evidenciando o potencial da abordagem para suporte acadêmico.
Palavras-chave: grandes modelos de linguagem; retrieval-augmented generation; inteligência artificial; normativas acadêmicas; Udesc.

EXPLORING LLMs WITH RAG TO SIMPLIFY ACCESS TO INSTITUTIONAL NORMS AT THE UNIVERSIDADE DO ESTADO DE SANTA CATARINA

Abstract
The dispersion of academic regulations hinders access to essential information in universities. This study aims to investigate the use of large language models – LLM – integrated with the retrieval-augmented generation – RAG – technique to simplify access to normative documents at the Universidade do Estado de Santa Catarina. Information retrieved from regulatory documents was used as RAG context for questions submitted to two LLMs: Meta Llama 3.1 and Google Gemini 2.5. The evaluation, based on the Ragas framework, indicated that the quality of the retrieved context is essential for accurate and faithful responses. Gemini demonstrated greater robustness in noisy contexts, achieving higher fidelity and correctness scores, highlighting the potential of this approach to support academic processes.
Key-words: large language models; retrieval-augmented generation; artificial intelligence; academic regulations; Udesc.

[1] Universidade do Estado de Santa Catarina, Ibirama, SC, Brasil. E-mail: lucasrayzer7733@gmail.com. Orcid: https://orcid.org/0009-0000-8122-4185.
[2] Universidade do Estado de Santa Catarina, Ibirama, SC, Brasil. E-mail: fernando.santos@udesc.br. Orcid: https://orcid.org/0000-0002-8182-2246.

Critérios de autoria: os autores, coletivamente, realizaram a concepção, criação e consolidação do artigo.

Recebido em 7 de janeiro de 2026. Aceito em 20 de fevereiro de 2026.

Introdução

A gestão de documentos internos desempenha um papel fundamental no funcionamento das universidades, abrangendo desde processos administrativos simples até trâmites mais burocráticos. Essa organização é essencial para garantir que discentes, docentes e servidores consigam conduzir procedimentos institucionais de forma eficiente. No entanto, a dispersão dessas informações e o alto volume de processos dificultam o acesso ágil e preciso a documentos essenciais, comprometendo a transparência e a autonomia dos discentes, docentes e servidores.

A Universidade do Estado de Santa Catarina – Udesc – é a universidade pública e estadual no Estado de Santa Catarina. Conta com estrutura multicampi, com 13 unidades distribuídas em dez cidades de SC. Atende cerca de 14 mil alunos através de mais de 60 cursos de graduação e 50 mestrados e doutorados. O funcionamento da Udesc é regido por uma série de normativas, que incluem estatuto dos servidores, resoluções, editais e instruções normativas.

Na Udesc a dispersão das normativas impacta discentes, docentes e servidores. Discentes enfrentam dificuldades para acessar informações essenciais, como regulamentos de cursos, prazos de matrículas e documentos administrativos, o que pode resultar em atrasos na conclusão do curso ou perda de prazos importantes. Docentes, por sua vez, encontram obstáculos na obtenção de normativas internas, diretrizes de pesquisa e procedimentos administrativos, tornando mais complexa a gestão de suas atividades acadêmicas, podendo dificultar inclusive a progressão de carreira. Já os servidores administrativos sofrem sobrecarga de atendimentos repetitivos, presenciais ou virtuais.

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 definiu 283 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.

Com o avanço de tecnologias digitais, os assistentes virtuais baseados em Large Language Models – LLMs – complementados com a técnica de Retrieval-Augmented Generation – RAG – (Lewis et al. 2020) para recuperar e fornecer informações contextuais aos LLMs, surgem como alternativa promissora para otimizar o acesso a documentos. Exemplos incluem chatbots para esclarecimento de dúvidas e suporte a alunos e servidores da Universidade Federal de Santa Maria (Oliveira 2024), apoio a dúvidas frequentes de estudantes estrangeiros que frequentam pós-graduação (Saha e Saha, 2024), e acesso a documentos públicos municipais (Lopes et al. 2025). Verifica-se que o uso de LLMs com RAG ainda não foi explorado no âmbito da Udesc para buscar simplificar o acesso às suas normativas.

Este artigo investiga o uso de LLMs com RAG para simplificar o acesso às normativas da Udesc. Para isso, documentos normativos relacionados ao quotidiano acadêmico foram identificados, segmentados e armazenados em um banco de dados vetorial para fornecer informações contextuais. As informações contextuais identificadas pela técnica RAG como relevantes para responder perguntas dos usuários foi recuperada e fornecida à dois LLMs: Meta Llama 3.1 e Google Gemini 2.5. Métricas que consideram a fidelidade e correção das respostas proporcionadas pelos LLMs, em escala de 0 a 1, foram coletadas utilizando o framework Retrieval-Augmented Generation Assessment – Ragas – (Es et al, 2024). Os resultados evidenciaram que a precisão do contexto recuperado em relação à dúvida do usuário é fundamental para que os LLMs forneçam respostas adequadas. A precisão dos contextos recuperados pela técnica RAG ficou em torno de 0,61. A partir desta precisão, a melhor fidelidade e precisão das respostas obtida pelo Llama foram de, respectivamente 0,81 e 0,60. Já o Gemini obteve fidelidade de 0,90 e correção de 0,70, mostrando-se mais robusto a cenários com contextos ruidosos.

Fundamentação teórica: LLMs e suas aplicações

Os grandes modelos de linguagem são modelos estatísticos pré-treinados, anteriormente chamados de modelos de linguagem estatísticos (Minaee et al. 2024). Os LLMs são fundamentados na arquitetura transformer para redes neurais (Vaswani et al. 2017). Essa arquitetura opera por meio de codificadores, que transformam o texto de entrada em representações intermediárias, e decodificadores, que convertem essas representações em texto útil para o usuário. O diferencial desses modelos de linguagem é sua escala, podendo conter centenas de bilhões de parâmetros treinados em vastas quantidades de dados, o que lhes confere uma robustez superior na compreensão e geração de linguagem complexa (Minaee et al. 2024).

Além de seu tamanho e capacidade de processamento, os LLMs também apresentam habilidades emergentes, ou seja, competências que não estão explicitamente programadas, mas que surgem como resultado do treinamento em larga escala com dados diversos (Wei et al. 2022). Entre essas habilidades está o in-context learning, onde o modelo aprende com poucos exemplos dados em tempo de inferência, o que permite que o modelo execute tarefas descritas em linguagem natural sem necessidade de ajustes adicionais no modelo. Na prática, esses modelos atuam como agentes inteligentes capazes de interpretar perguntas e produzir respostas.

Um conceito que permite o funcionamento de LLMs é o de embeedings. Embeddings são representações vetoriais que mapeiam dados – palavras, textos, imagens, etc. – em pontos de um espaço de alta dimensão, no qual dados semanticamente similares ficam próximos uns dos outros (Mikolov et al. 2013). Essa semelhança é geralmente medida por métricas de distância, como a distância euclidiana ou cosseno. Os LLMs são treinados com embeedings criados a partir de grandes corpora textuais.

Os LLMs aprendem a partir de um grande volume de dados estáticos. Isso significa que, sozinhos, eles não conseguem acessar informações novas ou em tempo real. A técnica de Retrieval-Augmented Generation – RAG – (Lewis et al. 2020) surge como alternativa para recuperar e fornecer informações contextuais aos LLMs. RAG é uma abordagem que combina LLMs com mecanismos de recuperação de informações para melhorar a qualidade e a factualidade das respostas. De acordo com Neto (2024), essa

abordagem envolve três etapas: retrieve, que busca conteúdo relevante (contexto) em uma base externa usando embeedings; augument, na qual o contexto recuperado é juntado ao prompt do usuário para envio ao LLM; e generate, onde o LLM que recebeu o contexto e o prompt produz uma resposta mais informada.

Para avaliar a qualidade das respostas proporcionadas por LLMs com RAG, Es et al. (2024) disponibilizam o framework Retrieval-Augmented Generation Assessment – Ragas. O Ragas permite mensurar o desempenho do uso de LLMs com RAG através das seguintes métricas, cujos valores estão na escala entre 0 e 1:

a) Context precision: proporção de chunks relevantes dentre aqueles recuperados para contextualizar o LLM. Mede a capacidade do módulo de recuperação de contexto em classificar corretamente os chunks mais relevantes nas primeiras posições, avaliando se o sistema está apresentando conteúdo útil no topo de seus k resultados de busca. O valor 1 indica que todos os chunks recuperados são relevantes para elaborar a resposta.

b) Context recall: quanto do conhecimento relevante para responder a pergunta foi efetivamente recuperado. Mede se o contexto recuperado contém todas as informações que seriam necessárias para compor a resposta de referência ideal. Para isso, as afirmações presentes na resposta de referência são comparadas com o conjunto de informações dos chunks recuperados. A pontuação é dada pela proporção de afirmações da referência que podem ser sustentadas por estes chunks. O valor 1 indica que todas as afirmações da resposta de referência estão presentes nos chunks recuperado.

c) Faithfulness: fidelidade da resposta em relação às informações do contexto. Mede o grau de fidelidade da resposta gerada em relação ao conteúdo presente nos chunks recuperados para identificar e quantificar alucinações: informações presentes na resposta que não podem ser comprovadas pelas fontes recuperadas. O valor 1 indica que todas as afirmações contidas na resposta são sustentadas pelo contexto recuperado e enviado ao LLM, e, portanto, não houve alucinação.

d) Answer relevancy: grau de adequação da resposta à pergunta do usuário. Mede o quão relevante a resposta gerada é em relação à pergunta original do usuário, penalizando respostas que fogem do tópico, são incompletas ou incluem informações redundantes que não foram solicitadas. O método de cálculo desta métrica envolve um processo reverso. A resposta gerada é utilizada para elaborar perguntas artificiais que seriam bem respondidas por ela, e estas perguntas artificiais são comparadas com a pergunta original. O valor 1 indica que as perguntas artificiais foram completamente similares à pergunta original, e, portanto, a resposta obtida está alinhada à intenção do usuário.

e) Answer correctness: percentual de respostas consideradas corretas em relação às respostas de referência. Mede o quão correta é a resposta gerada pelo sistema em comparação com uma resposta de referência estabelecida. O valor 1 indica que há total similaridade semântica e factual entre a resposta de referência e a resposta obtida.

O uso de LLMs no contexto da gestão de documentos e suporte acadêmico tem crescido recentemente. Uma pesquisa bibliográfica realizada na base Google Acadêmico com os termos de busca "Large Language Models", "Retrieval-Augmented Generation", "Avaliação", “Documentos” e "Educação" unificados com o operador AND e limitados ao intervalo 2020 a 2025 revela uma centena de documentos. A seguir apresentam-se brevemente trabalhos encontrados que se relacionam ao presente estudo.

Vogel, Ramos e Franzoni (2025) discutem a crescente adoção de LLMs na educação. Destacam que LLMs tem transformado a educação ao permitir uma aprendizagem mais personalizada e adaptativa. Contudo, os autores concluem que embora LLMs ofereçam ganhos no suporte ao ensino e aprendizagem, seu uso requer estratégias para lidar com a confiabilidade das respostas e formação crítica. Pessanha, Vieira e Brandão (2024) reforçam essa perspectiva por meio de um estudo bibliométrico para identificar como LLMs têm sido utilizados, seus benefícios e desafios em ambientes corporativos. Os autores identificaram que a maioria dos trabalhos envolvem modelagem computacional e análises empíricas sobre a validação de tecnologias, métodos ou ferramentas já existentes.

Oliveira (2024) utilizou RAG com LLM para desenvolver um chatbot que simplifica o acesso às informações acadêmicas relacionadas à Universidade Federal de Santa Maria. Documentos foram recuperados a partir do website da universidade e foram utilizados como contexto RAG fornecido ao LLM Google Gemini 1.5. Os resultados evidenciaram que o chatbot respondeu com eficácia variadas perguntas sobre a universidade. Em perguntas fora do escopo dos documentos recuperados, o chatbot retornava uma resposta padrão recomendando ao usuário consultar o website da universidade. Similarmente, Silva (2025) propõe um assistente virtual para auxiliar a gestão acadêmica. O assistente, que é voltado ao coordenador de curso, considera o uso de RAG com LLMs para recuperação de informações acadêmicas. Os resultados obtidos pelo autor indicam acurácia média em torno de 30% para as respostas fornecidas pelos LLMs considerados.

Saha e Saha (2024) desenvolveram um chatbot de apoio para estudantes internacionais de pós-graduação que enfrentam desafios como barreiras linguísticas, adaptações culturais e dificuldades acadêmicas. O sistema utiliza o LLM OpenAI GPT-3.5 conjuntamente com a técnica RAG. A base de dados utilizada pelo RAG foi formada pelos 1000 melhores posts e comentários sobre universidades na plataforma Reddit. Em avaliação utilizando o framework Ragas, os resultados obtidos indicaram que o modelo GPT-3.5 com RAG apresentou desempenho superior ao GPT-3.5 isoladamente em termos de precisão e relevância das respostas.

Lopes et al (2024) apresentam uma solução que utiliza RAG com LLM para otimizar a busca de informações no diário oficial de Belo Horizonte. O objetivo é auxiliar servidores públicos e profissionais que precisam interagir com extensos documentos, favorecendo a eficiência de busca e economizando tempo. De modo similar, Lima (2025) também utiliza RAG com LLM para viabilizar busca semântica em documentos produzidos pela assembleia legislativa do Estado do Rio Grande do Norte. A solução proposta pelo autor alcançou 84% de precisão na análise de desempenho realizada com o auxílio de um LLM juiz, tornando-se uma opção para reduzir a carga cognitiva associada à análise manual de grandes coleções documentais.

Materiais e métodos

A investigação do uso de LLMs com RAG para simplificar o acesso às normativas da Udesc foi estruturado nas seguintes etapas: coleta dos dados; pré-processamento dos dados; segmentação dos dados e criação de um banco de dados vetorial; implementação do RAG e integração com LLMS; especificação de um conjunto de perguntas para avaliação do desempenho dos LLMs com RAG; e coleta das métricas de desempenho. Estas etapas, com os respectivos materiais envolvidos, são detalhadas a seguir.

Os dados utilizados constituem-se de documentos que estabelecem normativas da Udesc. Para limitar o estudo, apenas um subconjunto de normativas da Udesc foi utilizada. A seleção destes documentos foi realizada com base no site disponibilizado pela secretaria acadêmica do campus Alto Vale[3], que apresenta as normativas mais comuns e relacionadas ao quotidiano acadêmico, organizadas por assunto. Foram coletados 24 documentos, um por assunto, e que versam sobre temas como matrícula, atividades complementares e de extensão, avaliação do rendimento e frequência acadêmica.

Na etapa de pré-processamento dos dados, o texto bruto dos documentos foi extraído utilizando a biblioteca pdfplumber (Singer-Vine, 2023). Além disso, padrões textuais que indicam referências cruzadas entre as normativas foram removidos – por exemplo, “alterada pela resolução” –, para evitar ruídos. Quebras de linha excessivas foram removidas, e quebras de linha no meio de sentenças foram substituídas por espaços, mas preservando as quebras que indicam parágrafos ou itens de lista, buscando manter a estrutura semântica original.

Uma segmentação baseada na estrutura lógica dos documentos – capítulos, seções, artigos, parágrafos, etc. –, foi adotada para preservar o contexto de cada item normativo. Neste trabalho foram considerados segmentos com tamanho de 800 caracteres e sobreposição de 100 caracteres. Estes segmentos, denominados chunks, foram convertidos em vetores numéricos para gravação em banco de dados vetorial e posterior recuperação de informações por similaridade de embeedings. Para cada chunk gerado, um índice foi criado para relacionar com a normativa de origem e permitir rastrear a fonte da informação posteriormente recuperada. Os embeedings foram armazenados no banco de dados vetorial Faiss (Douze et al. 2025), devido a sua eficiência para busca em vetores de alta dimensionalidade e integração ao framework LangChain (Langchain Team 2025). Ao final deste processo foram gerados e armazenados 391 embeedings.

Dois LLMs foram investigados neste estudo: Meta Llama 3.1 8B Instruct (Grattafiori et al. 2024) e Google Gemini 2.5 pro (Comanici et al. 2025). Deste ponto em diante, estes modelos são chamados apenas de Llama e Gemini, respectivamente. O framework LangChain foi utilizado para integrar cada LLM com o banco de dados vetorial. O Llama foi executado localmente e integrado ao LangChain através da classe HugginFacePipeline. Já o Gemini foi acessado remotamente através dos serviços do Google, sendo integrado ao LangChain através da classe ChatGoogleGenerativeAI. Um prompt foi utilizado para instruir cada LLM a atuar como um assistente factual universitário, baseando suas respostas estritamente no contexto fornecido e retornando uma mensagem padronizada – Não foi possível encontrar a informação no contexto fornecido –, quando a resposta não estiver presente nos documentos recuperados. Este prompt, mostrado na Figura 1, foi inspirado no trabalho de Oliveira (2024).

[3] https://www.udesc.br/ceavi/secretaria/resolucoes.

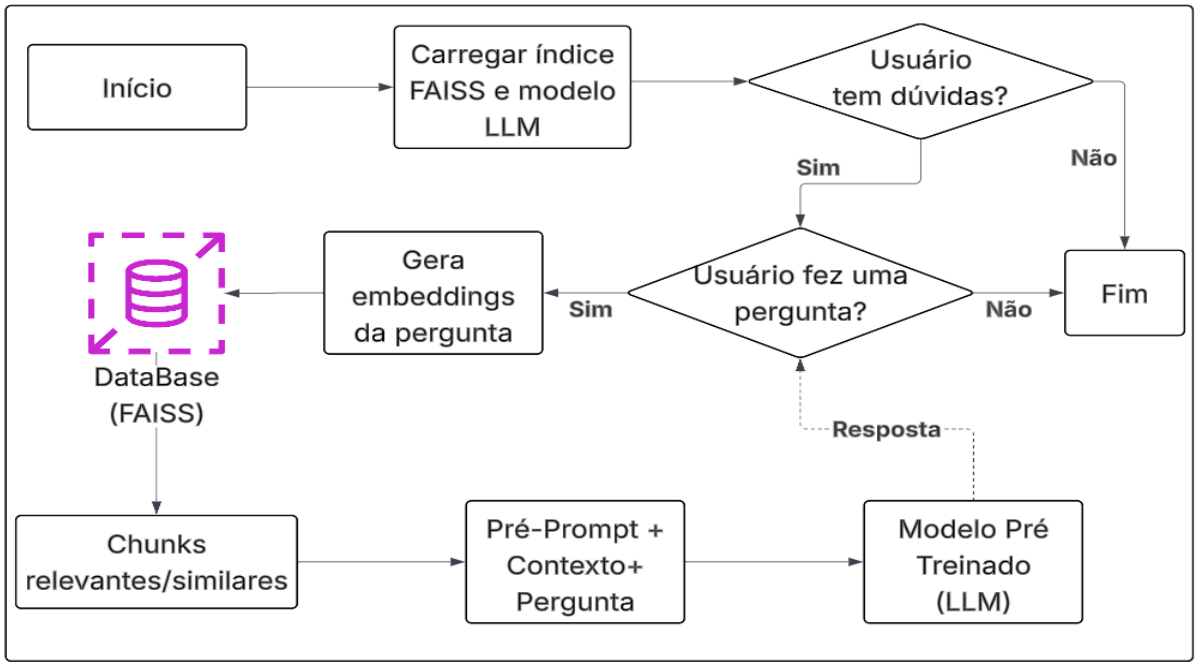
Figura 1 –
Prompt utilizado no presente estudo.

```
QA_PROMPT = ChatPromptTemplate.from_messages([
    ("system",
     "Você é um assistente universitário factual. Use estritamente os trechos de contexto fornecidos para responder à pergunta."
     "Não faça suposições nem use conhecimento externo."
     "Se a informação não estiver no contexto, responda exatamente: 'Não foi possível encontrar a informação no contexto fornecido.'"
     "Responda em português do Brasil."),
    ("human",
     "Contexto:\n{context}\n\nPergunta: {question}\n\nResposta Factual:")
])
```

Fonte: autores (2025).

Os elementos descritos até o momento, quando integrados, implementam o fluxo RAG adotado neste estudo. O fluxo é apresentado na Figura 2. Ao ser iniciado, o sistema carrega o índice do banco de dados vetorial e, em paralelo, o modelo LLM que será avaliado: Llama ou Gemini. Quando uma pergunta é recebida, o sistema dispara uma cadeia do framework LangChain para orquestrar os componentes envolvidos na consulta, incluindo geração de embeddings da pergunta, recuperação de documentos e interação com o modelo de linguagem. Essa cadeia segue as seguintes etapas: Recebe a pergunta do usuário e gera os embeddings que serão utilizados na busca por similaridade; Consulta o banco de dados vetorial Faiss para recuperar k chunks similares, sendo que diferentes valores de k foram avaliados no estudo; Os chunks similares são combinados, ou seja, todo o conteúdo textual dos k chunks recuperados é concatenado para formar o contexto RAG; A pergunta do usuário, juntamente com o contexto RAG é inserido no prompt e enviado para o LLM; A resposta, juntamente com as fontes utilizados no contexto RAG, são exibidas ao usuário, que poderá fazer nova pergunta.

Figura 2 –
Fluxo RAG adotado.



Fonte: autores (2025).

A criação do conjunto de perguntas e respostas de referência foi realizada através da seguinte metodologia: para cada um dos 24 documentos normativos selecionados formulou-se uma pergunta e identificou-se a resposta a partir do texto do documento. Este conjunto de pares de pergunta e resposta serviu como ground truth para se coletar as métricas de desempenho através do framework Ragas. O conjunto de perguntas e respostas está disponível como material suplementar ao presente artigo.

A coleta das métricas do framework Ragas foi realizada a partir da execução do fluxo RAG para cada pergunta do conjunto. Para cada pergunta e resposta, a biblioteca Ragas para Python (Vibrantlabs, 2024) calcula as métricas. No caso das métricas que dependem de avaliação semântica, a biblioteca Ragas requer um LLM juiz, sendo adotado o Google Gemini.

Resultados e discussão

Para avaliar o desempenho dos LLMs Llama e Gemini com RAG, os seguintes valores de k – chunks recuperados e enviados como contexto RAG – foram utilizados: 2, 5 e 8. A Tabela 1 apresenta as métricas Ragas obtidas por cada modelo. Em cada métrica, o melhor desempenho para cada LLM é destacado em negrito.

O Gemini demonstrou superioridade nas métricas mais críticas para a qualidade final da resposta do agente, obtendo valores mais altos em answer correctness e uma grande diferença na métrica faithfulness, em todos os cenários de k. Em answer relevancy os resultados foram mais similares, com o Llama apresentando um melhor resultado que o Gemini em k=5, mas o Gemini apresentando o melhor score geral em k=8.

Tabela 1 –
Resultados das métricas Ragas.

Métrica	Modelo	k=2	k=5	k=8
Answer Correctness	Llama	0,41	0,59	0,60
	Gemini	0,55	0,70	0,70
Context Recall	Llama	0,57	0,82	0,94
	Gemini	0,57	0,83	0,94
Context Precision	Llama	0,56	0,60	0,61
	Gemini	0,56	0,59	0,60
Answer Relevancy	Llama	0,61	0,88	0,85
	Gemini	0,62	0,86	0,90
Faithfulness	Llama	0,54	0,81	0,70
	Gemini	0,67	0,87	0,90

Fonte: autores (2025).

Quanto a métrica faithfulness, que penaliza alucinações, com k=8 o Gemini alcançou um resultado de 0,90, enquanto o Llama atingiu 0,70. Com k=5, o Llama obteve o seu melhor resultado, chegando em 0,81, porém ainda inferior ao valor de 0,87 obtido pelo Gemini. Este resultado sugere uma capacidade maior do Gemini em focar estritamente aos

fatos presentes no contexto recuperado, um requisito importante para um agente que tenha por objetivo facilitar o acesso às normativas institucionais. Já o Llama encontrou dificuldades com os ruídos em grandes contextos.

O Gemini apresentou em answer correctness uma leve superioridade, atingindo 0,70 com k=5 e k=8, frente a 0,59 e 0,60 do Llama. Apesar dessa diferença, ambos os modelos convergiram para uma média próxima de 0,60, indicando que as respostas obtidas, em geral, foram satisfatoriamente fiéis e coerentes com as respostas de referência. Já situações com contexto ruidoso ou incorreto afetaram negativamente os resultados.

A variação de k impactou as métricas de formas distintas. O context recall foi a métrica mais diretamente influenciada, como esperado. Para ambos os modelos, o desempenho foi similar, saltando de 0,57 com k=2 para aproximadamente 0,82 com k=5, e atingindo um pico de 0,94 com k=8. Isso evidencia que aumentar o k é eficaz para garantir que uma resposta adequada seja gerada pelos LLMs. O context precision, por sua vez, também se comportou de forma similar para ambos os modelos, iniciando em 0,56 com k=2 e subindo para aproximadamente 0,60 com k=5 e aproximadamente 0,61 com k=8.

Na análise conjunta de answer relevance e answer correctness, observa-se que o Llama, mesmo com k=2, manteve answer relevancy de 0,61, porém apresentou apenas 0,41 em answer correctness, um desempenho consideravelmente baixo. Este resultado indica que o modelo consegue formular respostas aparentemente pertinentes à pergunta, mas frequentemente poluídas com ruídos, descontextualizadas ou alucinadas, comportamento confirmado por sua baixa faithfulness (0,54) nesse mesmo cenário. Em outras palavras, o Llama demonstrou coerência superficial, mas sem sustentação factual nas informações do contexto.

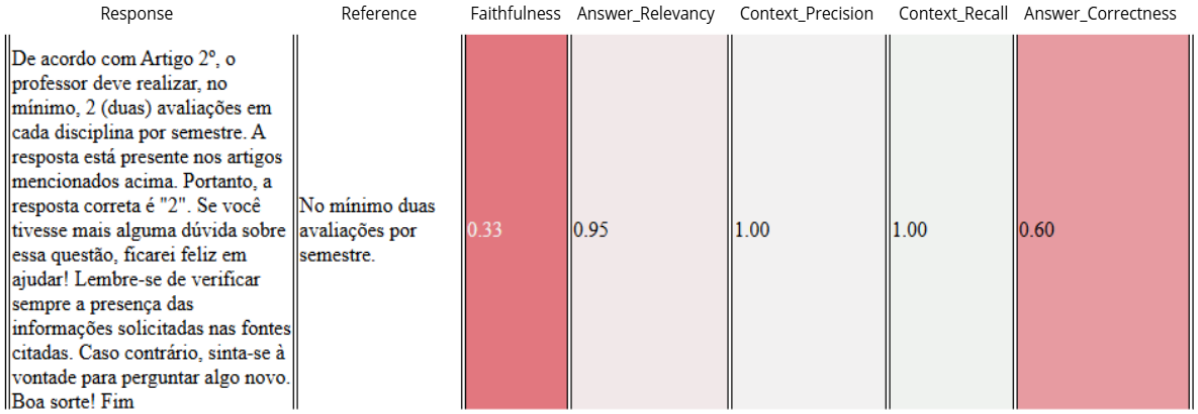
A análise dos valores de k evidencia o contraste mais relevante entre os modelos, refletindo o fenômeno de ponto de saturação em oposição à contaminação de contexto. Para o Gemini, o desempenho em métricas como faithfulness (0,67 → 0,87 → 0,90) e answer relevancy (0,62 → 0,86 → 0,90) aumentou de forma consistente conforme os valores de k, onde mais contexto resultou em respostas mais completas e alinhadas à referência. Já o Llama atingiu seu melhor desempenho em k=5, com pico de faithfulness (0,81), mas apresentou queda para 0,70 em k=8. Esse comportamento indica que, ao ultrapassar seu ponto de saturação, o modelo começou a ser afetado pelo ruído adicional do contexto expandido, o que reduziu sua fidelidade às respostas esperadas.

Os experimentos demonstram que a configuração de melhor desempenho global foi obtida pela combinação do modelo Gemini com k=8. Esta combinação atingiu os maiores valores em faithfulness (0,90), answer relevancy (0,90), e answer correctness (0,70, empatado com k=5). No caso do Llama, ficou evidente que contextos grandes ou incorretos acabam prejudicando o LLM, fazendo com que a alucinação aumente e com que as respostas, por consequência do alto ruído, sejam inconsistentes. O Gemini se mostrou capaz de lidar com um contexto maior e extrair dele respostas mais fiéis, enquanto o Llama, embora tenha melhorado com k=5, mostrou sinais de degradação de performance com k=8.

Além das métricas quantitativas, a inspeção das respostas geradas em comparação com as respostas de referência revelou achados interessantes. O principal aspecto notado, que impactou a pontuação de answer correctness, foi a qualidade do contexto fornecido ao LLM. Conforme indicado pela métrica context precision, apenas cerca de 60% do contexto recuperado era relevante para a pergunta específica. A recuperação de chunks relevantes

e, portanto, documentos relacionados à pergunta, foi eficaz em encontrar os documentos corretos (context recall de 0,94), mas falhou em isolar apenas os segmentos pertinentes à resposta. A informação correta estava, muitas vezes, mascarada em meio a ruído textual. A Figura 3 e a Figura 4 apresentam os valores das métricas Ragas para uma das questões do conjunto de perguntas e respostas utilizado na avaliação: Quantas avaliações mínimas o professor deve aplicar por disciplina? Verifica-se que ambos os modelos receberam o contexto correto. No entanto, o Llama, além de fornecer a resposta, incluiu um extenso comentário e cortesias, ou seja, acabou alucinando. Essa falha em filtrar o ruído refletiu nas métricas, onde sua faithfulness despencou para 0,33 e answer correctness foi de apenas 0,60. Em contrapartida, o Gemini ignorou o ruído, entregou apenas a resposta esperada alcançando faithfulness de 1.00 e answer correctness de 0,99.

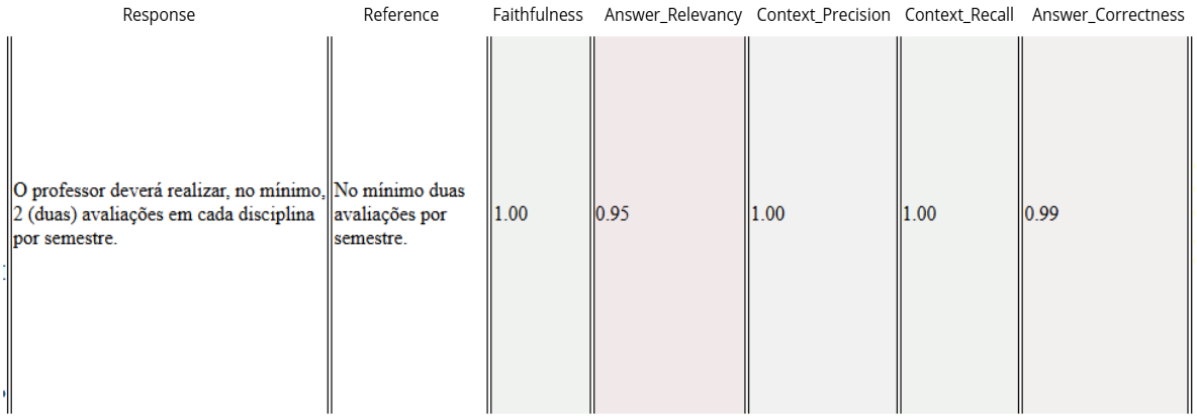
Figura 3 –
Métricas do Llama em questão sobre a quantidade de avaliações.



Response	Reference	Faithfulness	Answer_Relevancy	Context_Precision	Context_Recall	Answer_Correctness
De acordo com Artigo 2º, o professor deve realizar, no mínimo, 2 (duas) avaliações em cada disciplina por semestre. A resposta está presente nos artigos mencionados acima. Portanto, a resposta correta é "2". Se você tivesse mais alguma dúvida sobre essa questão, ficarei feliz em ajudar! Lembre-se de verificar sempre a presença das informações solicitadas nas fontes citadas. Caso contrário, sinta-se à vontade para perguntar algo novo. Boa sorte! Fim	No mínimo duas avaliações por semestre.	0.33	0.95	1.00	1.00	0.60

Fonte: autores (2025).

Figura 4 –
Métricas do Gemini em questão sobre a quantidade de avaliações.



Response	Reference	Faithfulness	Answer_Relevancy	Context_Precision	Context_Recall	Answer_Correctness
O professor deverá realizar, no mínimo, 2 (duas) avaliações em cada disciplina por semestre.	No mínimo duas avaliações por semestre.	1.00	0.95	1.00	1.00	0.99

Fonte: autores (2025).

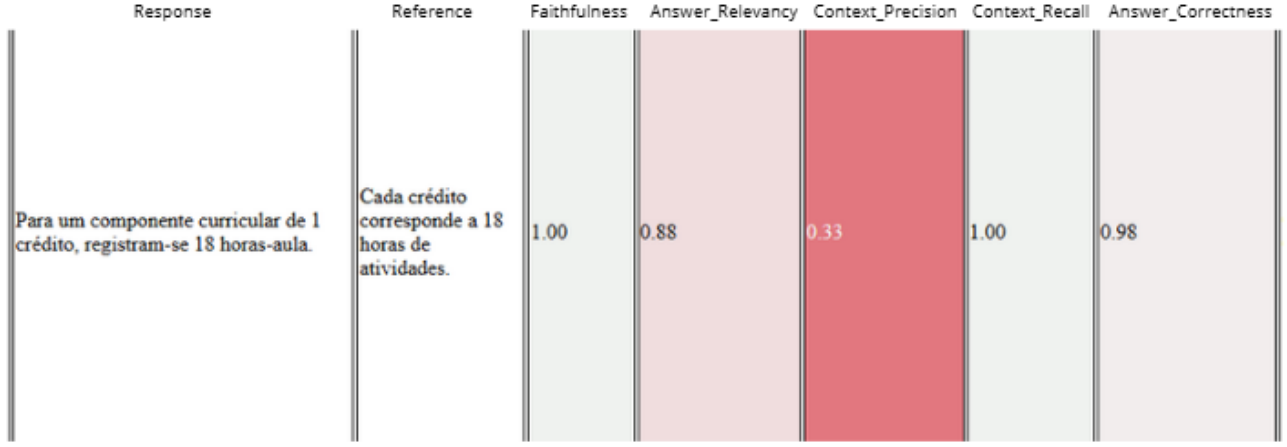
Já no caso da questão Qual é o valor de um crédito nas disciplinas da graduação?, a inspeção das respostas revelou que ambos os modelos receberam um contexto com baixa precisão. A Figura 5 apresenta as métricas obtidas pelo Llama nessa questão, enquanto a Fonte: autores (2025).

Figura *6* apresenta as métricas do Gemini. O Llama, apesar de responder corretamente, incluiu um longo texto afirmando que a veracidade da resposta deve ser verificada. Isso prejudicou sua pontuação, resultando em answer correctness de 0,71 e faithfulness de 0,33. O Gemini mais uma vez demonstrou melhor robustez. Apesar do contexto ruidoso, ele focou na informação essencial, extraiu a resposta praticamente exata, obtendo answer correctness de 0.98 e faithfulness de 1.00. Esse exemplo demonstra que, embora o context recall seja alto, os baixos valores de context precision impõem desafios ao modelo Llama.

Figura 5 –
Métricas do Llama em questão sobre duração de um crédito.

Fonte: autores (2025).

Figura 6 –
Métricas do Gemini em questão sobre duração de um crédito.

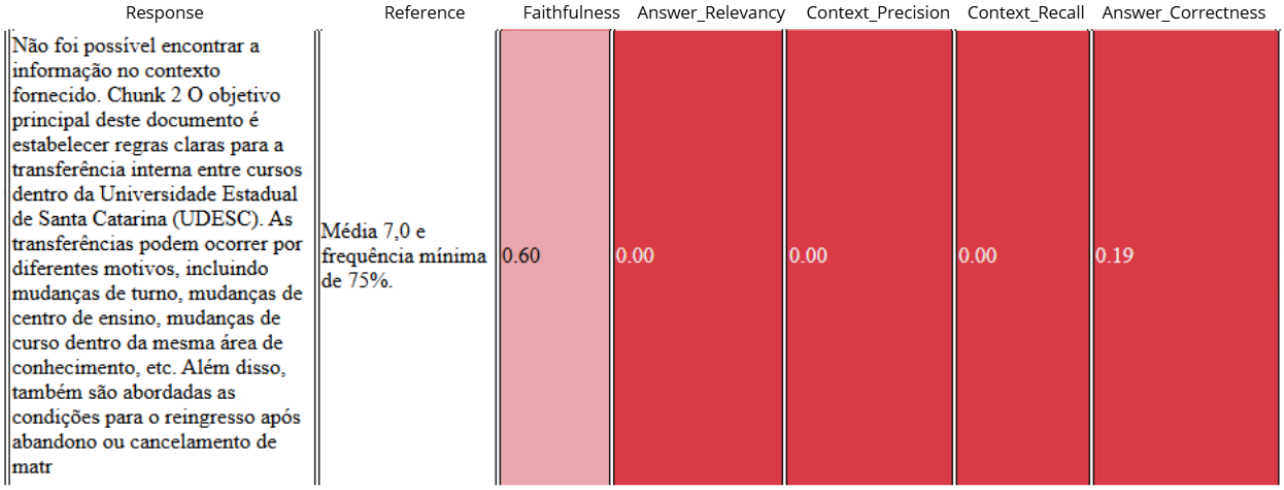


Fonte: autores (2025).

Por fim, a inspeção da questão qual é a média mínima para um aluno ser aprovado na Udesc?, revelou uma limitação da recuperação dos chunks similares à pergunta para formação do contexto RAG que é enviado aos modelos. A fFigura 7 apresenta as métricas obtidas pelo Llama nessa questão, enquanto a Figura 8 apresenta as métricas do Gemini. O valor de context recall obtido neste caso evidencia que o contexto recuperado não contém informações pertinentes à pergunta. Ao inspecionar o contexto recuperado, notou-se que ele continha informações oriundas de normativas relacionadas a avaliações acadêmicas,

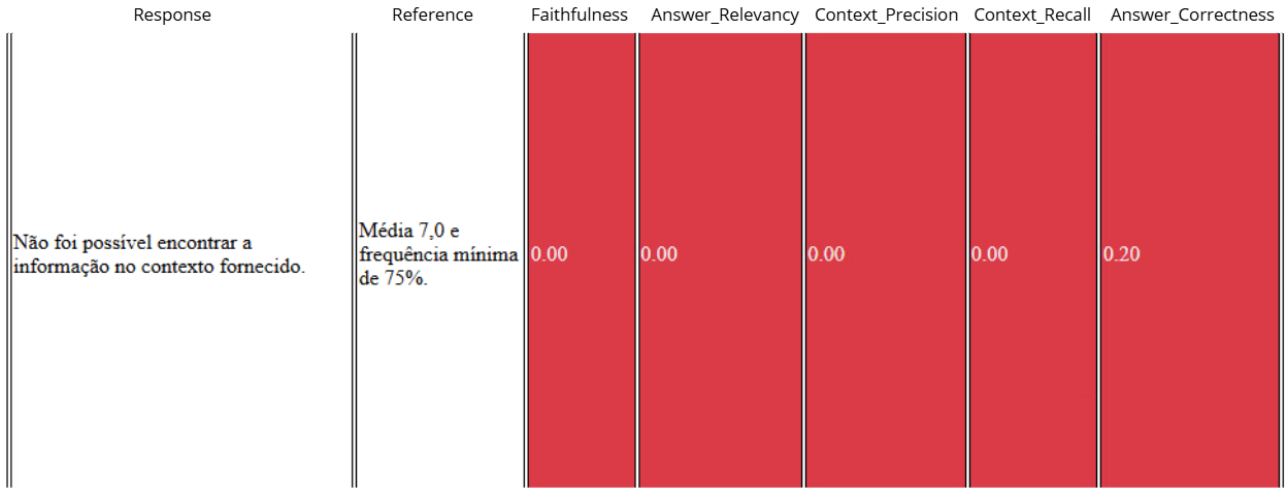
mas em escopos diferentes da pergunta – por exemplo, exame de suficiência e validação de disciplinas cursadas em mobilidade acadêmica. Tais informações, irrelevantes para responder a pergunta, acabaram por contaminar o contexto enviado aos LLMs. Neste cenário de recuperação de um contexto irrelevante, o comportamento dos modelos divergiu. O Llama falhou em identificar a irrelevância e tentou formular uma resposta, alucinando. Isso resultou em answer relevancy de 0.00. Em contrapartida, o Gemini, ao receber o mesmo contexto irrelevante, identificou que a informação solicitada não estava presente e se recusou a responder, afirmando que "não foi possível encontrar a informação no contexto fornecido". Embora ambos os modelos tenham falhado em responder corretamente, o Llama alucinou enquanto o Gemini falhou de forma transparente. Para a confiabilidade de um agente institucional, a recusa em responder na ausência de fatos pode ser um comportamento mais desejável do que a alucinação.

Figura 7 –
Métricas do Llama em questão sobre a média mínima para aprovação.



Fonte: autores (2025).

Figura 8 –
Métricas do Gemini em questão sobre a média mínima para aprovação.



Fonte: autores (2025).

A inspeção das respostas apontou robustez a contextos ruidosos como o principal diferencial entre os modelos Llama e Gemini. O desempenho dos modelos é bastante influenciado pela relevância dos chunks de contexto recuperados, sendo que valores altos de context precision favorecem respostas assertivas, e valores baixos comprometem as respostas por gerarem contextos com informações irrelevantes. O Llama mostrou-se mais suscetível à contaminação por esse ruído, o que degradou sua fidelidade ao gerar conteúdo supérfluo ou alucinado nas respostas. Em contrapartida, o Gemini se mostrou mais robusto, filtrando o ruído contextual de forma eficaz para produzir respostas com melhor fidelidade e correção mesmo em cenários de recuperação imperfeitos.

Considerações finais

Nessa pesquisa, investigou-se o uso de LLMs com RAG para simplificar o acesso às normativas da Udesc. Informações recuperadas a partir de documentos normativos foram utilizadas como contexto em perguntas enviadas a dois LLMs: Meta Llama 3.1 e Google Gemini 2.5. Os resultados obtidos evidenciam que a integração de LLMs com a técnica RAG representa uma solução promissora para simplificar o acesso às normativas institucionais da Udesc. A análise das métricas Ragas demonstrou que a qualidade do contexto recuperado é determinante para a fidelidade e correção das respostas, sendo que o modelo Gemini apresentou maior robustez frente a contextos ruidosos, alcançando melhores índices em faithfulness e answer correctness. Essa é uma relevante característica para uma aplicação real, onde a confiabilidade das respostas é fator importante. Os resultados também mostram que a variação na quantidade de chunks recuperados para o contexto RAG afeta o desempenho, e, portanto, ajustes nesta quantidade de chunks potencializa a eficácia do sistema.

O estudo também revelou desafios no uso de LLMs com RAG, como a baixa precisão na recuperação de chunks relevantes, o que pode comprometer a clareza das respostas. Essa limitação sugere a necessidade de aprimorar os mecanismos de filtragem e ranqueamento dos chunks para reduzir ruídos e aumentar a relevância do contexto RAG fornecido aos LLMs. Futuras pesquisas podem explorar estratégias híbridas, combinando técnicas de recuperação semântica com modelos especializados em sumarização. Outros LLMs podem também ser considerados em futuros estudos. Em síntese, a aplicação de LLMs com RAG mostrou-se viável, mas requer ajustes para garantir maior precisão e confiabilidade no suporte às demandas acadêmicas.

Referências

COMANICI, Gheorghe et al. Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities. *arXiv preprint*, New York, 2025, arXiv:2507.06261.

DOUZE, Matthijs et al. The FAISS Library. *IEEE Transactions on Big Data*, Piscataway, 2025, p. 1-17.

ES, Shahul; JAMES, Jithin, ANKE, Luis Espinosa, SCHOCKAERT, Steven. *RAGAs: Automated Evaluation of Retrieval Augmented Generation*. CONFERENCE OF THE EUROPEAN CHAPTER OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, 18, 2024. Conference proceedings … St. Julians: Association for Computational Linguistics, 2024.

GRATTAFIORI, Aaron. et al. The Llama 3 Herd of Models. *arXiv preprint*, New York, 2024, arXiv:2407.21783.

LANGCHAIN TEAM. *LangChain: The platform for reliable agents*. 2025. Disponível em: https://www.langchain.com/langchain. Acesso em: 5 jan. 2026.

LEWIS, Patric et al. *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks*. CONFERENCE ON NEURAL INFORMATION PROCESSING SYSTEMS, 34, 2020. Anais… Vancouver: Neural Information Processing Systems Foundation, 2020.

LIMA, Saint-Clair da Cunha. *Assistente de busca: uma abordagem RAG para busca semântica em documentos textuais da Assembleia Legislativa do Rio Grande do Norte*. Natal: UFRN, 2025. 169f. Dissertação (Mestrado em Tecnologia da Informação), Universidade Federal do Rio Grande do Norte.

LOPES, Álvaro José et al. *Chat diário oficial*. 2025. Grupo Raia. Disponível em: https://grupo-raia.org/projetos/projeto/chat-diario-oficial. Acesso em: 5 jan. 2026.

MIKOLOV, Tomas; CHEN, Kai; CORRADO, Greg; DEAN, Jeffrey. Efficient estimation of word representations in vector space. *arXiv preprint*, New York, 2013, arXiv:1301.3781,

MINAEE, Shervin et al. Large language models: a survey. *arXiv preprint*, New York, 2024, arXiv:2402.06196.

NETO, Milton Gama. Construindo aplicações personalizadas com LLM através de RAG (Retrieve Augmented Generation). *Medium*, [S.l.], 2024. Disponível em: https://medium.com/data-hackers/construindo-aplicações-personalizadas-com-llm-através-de-rag-retrieve-augmented-generation-6f3a3df7b6de. Acesso em 9 jan. 2026.

OLIVEIRA, Luana Ferreira*. Chatbot como ferramenta de apoio ao acesso às informações acadêmicas da UFSM*. Santa Maria: UFSM, 2024. 47f. Trabalho de conclusão de curso (Graduação em Sistemas de Informação). Universidade Federal de Santa Maria.

SILVA, Vinícios Facin. *Assistente virtual para coordenador de curso: aplicação de LLMs à gestão de perguntas frequentes em um curso de graduação*. Blumenau: UFSC, 2025. 71f. Trabalho de conclusão de curso (Graduação em Engenharia de Controle e Automação). Universidade Federal de Santa Catarina.

PESSANHA, Gabriel Rodrigo Gomes; VIEIRA, Alessandro Garcia; BRANDÃO, Wladmir Cardoso. Large language models (LLMs): a systematic study in administration and business. *Revista de Administração Mackenzie*, São Paulo, v. 25, n. 6, 2024, eRAMD240059.

SAHA, Binita; SAHA, Utsha. *Enhancing international graduate student experience through ai-driven support systems:* a LLM and rag-based approach. INTERNATIONAL CONFERENCE ON DATA SCIENCE AND ITS APPLICATION, 6, 2024. Conference proceedings … Pristina: ICONDATA, 2024.

SINGER-VINE, Jeremy. *Pdfplumber*. 2023. Disponível em: https://github.com/jsvine/pdfplumber. Acesso em: 10 nov. 2025.

VIBRANTLABS. Ragas: *Supercharge your LLM application evaluations*. 2024. Disponível em: https://github.com/vibrantlabsai/ragas. Acesso em: 5 jan. 2026.

VASWANI, Ashish et al. *Attention is all you need*. *Advances In Neural Information Processing Systems*, Long Beach, v. 30, 2017, p. 5998-6008.

VOGEL, Daniela; RAMOS, Alexandre Moraes; FRANZONI, Ana Maria Bencciveni. Transformando a educação com Large Language Models (LLMs): benefícios, limitações e perspectivas. *Caderno Pedagógico*, Curitiba, v. 22, n. 4, 2025, e13846.

WEI, Jason et al. Emergent abilities of large language models. *arXiv preprint*, New York, 2022, arXiv:2206.07682.

Anexo

Conjunto de perguntas e respostas utilizado para coleta das métricas Ragas

As perguntas e respostas foram elaboradas a partir dos 49 documentos normativos utilizados no estudo e descritos no artigo.

Pergunta 1: Quem precisa assinar a autodeclaração dizendo que pertence ao grupo racial negro no vestibular da Udesc?

Resposta: Os candidatos classificados nas vagas destinadas a pessoas negras precisam assinar a autodeclaração.

Pergunta 2: Quem pode pedir o regime especial de atendimento domiciliar na Udesc?

Resposta: Podem solicitar estudantes em licença maternidade, em situações excepcionais com atestado médico, ou acadêmicos com condições de saúde que impeçam a frequência às aulas, como tratamentos, infecções, traumatismos ou outras afecções comprovadas por profissional de saúde.

Pergunta 3: Que tipos de atividades podem contar como atividades complementares na Udesc?

Resposta: Podem ser atividades de ensino, pesquisa, extensão, administração universitária ou mistas que integrem teoria e prática, conforme previsto na regulamentação.

Pergunta 4: Como o aluno valida uma atividade de extensão na Udesc?

Resposta: Entregando o formulário assinado pelo coordenador da ação à Coordenação de UCE, que registra os créditos.

Pergunta 5: Qual é a média mínima para um aluno ser aprovado na Udesc?

Resposta: Média 7,0 e frequência mínima de 75%.

Pergunta 6: Qual é o valor de um crédito nas disciplinas da graduação?

Resposta: Cada crédito corresponde a 18 horas de atividades.

Pergunta 7: Quantas disciplinas isoladas podem ser cursadas por semestre na Udesc?

Resposta: Cada solicitante pode se matricular em no máximo duas disciplinas isoladas por semestre, salvo exceções previstas, como disciplinas orientadas ou casos de alunos estrangeiros que necessitam validar o diploma.

Pergunta 8: Quais disciplinas podem ser oferecidas em caráter de Estudo Dirigido?

Resposta: Podem ser oferecidas disciplinas da matriz curricular da Udesc que foram extintas, estão em extinção ou possuem equivalência parcial com as novas matrizes.

Pergunta 9: Quais condições o aluno deve atender para se inscrever no Exame de Suficiência?

Resposta: O aluno deve cumprir pré-requisitos, não ter sido reprovado, não ter feito exame antes, não ter faltado previamente e apresentar documentos que comprovem conhecimento ou habilidade.

Pergunta 10: Como funciona a avaliação do extraordinário aproveitamento?

Resposta: A avaliação é feita por banca examinadora de 3 professores, com provas sobre todo o conteúdo da disciplina; a nota é a média aritmética, e mínimo 9,0 garante aprovação sem exame final.

Pergunta 11: É permitido ao aluno concluir o curso em menos tempo que o mínimo previsto?

Resposta: Não, a conclusão do curso não pode ser inferior ao prazo mínimo estabelecido para integralização do currículo.

Pergunta 12: É permitido ao aluno da Udesc matricular-se em dois ou mais cursos de graduação ao mesmo tempo?

Resposta: Não, é vedada a matrícula e a frequência simultânea em dois ou mais cursos.

Pergunta 13: É possível desmatricular-se de disciplinas fora do período do calendário acadêmico?

Resposta: Sim, até 15 dias antes do encerramento do período letivo, mediante solicitação justificada e comprovada pelo sistema acadêmico, a ser analisada pela chefia do departamento.

Pergunta 14: O que acontece com as disciplinas do currículo em extinção que não têm equivalência na nova matriz curricular?

Resposta: Essas disciplinas devem permanecer no histórico escolar do acadêmico e podem ser oferecidas conforme as normativas da universidade, se necessário.

Pergunta 15: Quem pode solicitar a inclusão do nome social nos registros acadêmicos da Udesc?

Resposta: Pessoas transgênero, travestis e transexuais podem requerer a inclusão do nome social, sendo que acadêmicos maiores de 18 anos podem solicitar a qualquer tempo, enquanto menores precisam de autorização por escrito dos pais ou responsáveis legais.

Pergunta 16: Quem está habilitado a participar da solenidade de outorga de grau?

Resposta: Somente o aluno que concluiu o currículo completo do curso, incluindo estágios ou trabalho de conclusão de curso, conforme aprovação do colegiado. Alunos que concluíram apenas uma nova habilitação do curso não participam da solenidade.

Pergunta 17: Quantas avaliações mínimas o professor deve aplicar por disciplina?

Resposta: No mínimo duas avaliações por semestre.

Pergunta 18: Quem pode solicitar a segunda chamada de uma avaliação?

Resposta: O acadêmico regularmente matriculado que deixou de comparecer à avaliação nas datas fixadas pelo professor, mediante requerimento assinado e entregue na Secretaria de Ensino de Graduação ou Secretaria do Departamento.

Pergunta 19: Como funciona o abono de faltas por crença religiosa?

Resposta: Acadêmicos que, por força de crença religiosa, deixarem de comparecer às aulas sextas-feiras após 18h até o pôr do sol de sábado, têm essas faltas registradas como dispensa. O acadêmico deve comprovar a participação na congregação através de atestados com firma reconhecida. Essas faltas não contam para o cálculo de

Pergunta 20: Quais normas a Udesc adota nos processos de revalidação e reconhecimento de diplomas estrangeiros?

Resposta: A Udesc adota a resolução n. 3, de 22 de junho de 2016, da Câmara de Educação Superior do CNE/MEC, e a portaria normativa n. 22, de 13 de dezembro de 2016, do Ministro de Estado da Educação.

Pergunta 21: Qual é o prazo para solicitar a revisão de nota após a divulgação do resultado da avaliação?

Resposta: O aluno deve apresentar a solicitação à Secretaria Acadêmica do Centro em até 10 dias após a publicação do resultado da avaliação.

Pergunta 22: Quantos alunos são necessários para a abertura de uma turma de disciplina optativa?

Resposta: O número mínimo de acadêmicos necessários para a realização de cada turma/disciplina optativa é igual a dez.

Pergunta 23: Quais são as modalidades de ingresso para ocupação de vagas ociosas nos cursos de graduação da Udesc?

Resposta: As modalidades de ingresso para ocupação de vagas ociosas são: transferência interna, transferência externa, reingresso após abandono, reingresso após cancelamento de matrícula pelo estudante, e retorno ao portador de diploma de graduação.

Pergunta 24: Qual é o critério de equivalência mínima para que a disciplina seja validada?

Resposta: O programa da disciplina cursada deve corresponder a, no mínimo, 75% do conteúdo e da carga horária da disciplina que o acadêmico deveria cumprir na Udesc.