Avaliação de um chatbot acadêmico com Ragas: Um estudo no IFPR Campus Paranavaí

Isaque Dias Prates da Cruz[1], **Frank Willian Cardoso de Oliveira**[1]

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

20230006245@estudantes.ifpr.edu.br, frank.willian@ifpr.edu.br

1. Introdução

Avaliar sistemas de inteligência artificial generativa impõe desafios que vão além da verificação de uma saída previamente definida. Modelos de linguagem podem apresentar comportamento não determinístico, produzindo respostas distintas para uma mesma entrada. Ademais, diferentes formulações podem expressar uma informação correta, enquanto respostas fundamentadas podem conter erros ou não atender à solicitação. Dessa forma, a comparação literal com uma resposta esperada é insuficiente para examinar a qualidade desses sistemas.

Nos sistemas de geração aumentada por recuperação, conhecidos como RAG (*Retrieval-Augmented Generation*), esse desafio envolve tanto a busca de informações quanto sua utilização na resposta. O Ragas oferece métricas para examinar essas dimensões separadamente, permitindo investigar a relevância dos contextos, a fundamentação das afirmações e a qualidade das respostas [Es et al. 2024]. A ferramenta não elimina a variabilidade dos modelos, mas fornece critérios para organizar a avaliação.

Nesse contexto, o presente trabalho apresenta a aplicação do Ragas na avaliação de um *chatbot* acadêmico desenvolvido para consultar documentos públicos do Instituto Federal do Paraná, Campus Paranavaí. O objetivo é examinar a recuperação e a geração de respostas, identificando limitações que orientem o aprimoramento do sistema.

2. Objeto de avaliação

O objeto de avaliação é um *chatbot* acadêmico desenvolvido para responder a dúvidas sobre o IFPR Campus Paranavaí com base em documentos institucionais. O sistema utiliza uma arquitetura RAG, cuja base de conhecimento reúne páginas web e arquivos PDF previamente processados e divididos em trechos, com representações vetoriais armazenadas no PostgreSQL com a extensão pgvector.

Ao receber uma dúvida acadêmica, o sistema produz uma consulta de busca a partir da pergunta e recupera trechos por similaridade vetorial. Os candidatos são reordenados por relevância e associados aos trechos maiores dos quais foram extraídos, conforme a estratégia de *Parent RAG*. Esses contextos são encaminhados ao modelo de linguagem juntamente com a pergunta para gerar a resposta. O modelo é orientado a utilizar exclusivamente as informações recuperadas e a indicar quando elas forem insuficientes para responder.

3. Procedimento de avaliação

O conjunto de avaliação foi composto por 17 fontes institucionais, sendo oito páginas web e nove PDFs, abrangendo informações sobre o curso de Engenharia de Software,

calendário acadêmico, biblioteca, estágio e serviços institucionais. Foram elaboradas 34 perguntas objetivas, duas por fonte, com auxílio de inteligência artificial. Os gabaritos foram produzidos a partir dos documentos, com revisão humana de todo o conjunto.

O sistema avaliado utilizou o modelo `gpt-5.4-nano` para geração e `text-embedding-3-small` para representação vetorial. A recuperação combinou reescrita da consulta com uma adaptação da técnica de *prompt* Hypothetic Document Embedding (HyDE), busca no PostgreSQL/pgvector e reordenação dos candidatos pelo modelo `gpt-5-nano`. O HyDE tornou-se essencial após o crescimento da base de dados, pois as perguntas em linguagem natural são geralmente curtas e coloquiais e os documentos têm linguagem formal, por isso foi mantido na avaliação. Já a estratégia de *Parent RAG* permitiu buscar trechos menores e entregar ao gerador os trechos maiores aos quais estavam associados.

Cada pergunta foi enviada à aplicação local em uma nova conversa, sem fornecer o gabarito e sem restringir a busca ao documento utilizado em sua elaboração. Foram coletadas a resposta final persistida e a totalidade dos contextos disponibilizados ao gerador, incluindo aqueles não citados na resposta. Esse procedimento evitou interferências do histórico e permitiu avaliar o contexto efetivamente utilizado pelo sistema.

A avaliação empregou o Ragas 0.4.3, com `gpt-4o-mini` como juiz nas etapas baseadas em modelos de linguagem e `text-embedding-3-small` nas comparações vetoriais de *Answer Relevancy*. Foram selecionadas cinco métricas, cujas finalidades são sintetizadas na Tabela 1, conforme a documentação da ferramenta [Ragas 2023]. Cada métrica foi calculada sobre as 34 questões, totalizando 170 pontuações.

4. Resultados e discussão

Como as métricas examinam aspectos distintos, não foi produzida uma média única entre elas. Também não foi estabelecido previamente um limiar de aceitação, de modo que a análise descreve o comportamento observado sem classificar o sistema como aprovado ou reprovado. A Tabela 1 apresenta as médias aritméticas das notas.

Tabela 1. Dimensões avaliadas e médias obtidas nas 34 questões

Métrica	Dimensão examinada	Média
Context Precision	Relevância e ordenação dos contextos em relação ao gabarito	0,664
Context Recall	Cobertura das informações do gabarito pelos contextos	0,828
Answer Relevancy	Correspondência da resposta com a pergunta	0,696
Faithfulness	Sustentação das afirmações nos contextos	0,900
Answer Accuracy	Concordância da resposta com o gabarito	0,544

A *Context Precision* foi inferior a 0,5 em 11 questões, em cinco delas o *Context Recall* foi 1,000, ou seja, o conteúdo foi recuperado, mas misturado a trechos pouco úteis ou posicionado após eles. No tema biblioteca, a Q17 (prazo de empréstimo) teve precisão e cobertura nulas, enquanto a Q18 (limite de renovações de um empréstimo), da mesma fonte, teve cobertura 1,000, sugerindo falha na recuperação do trecho, e não ausência da fonte.

A *Answer Relevancy* foi nula em cinco questões, e como o Ragas atribui zero a respostas vagas ou ambíguas, o padrão é compatível com penalização por falta de contexto. Dessas questões, quatro delas tiveram *Context Recall* inferior a 1, ou seja a recuperação não trouxe todos os dados esperados. A exceção é a Q03 (duração do curso de Engenharia de Software), com *Context Recall* de 1,000 e *Faithfulness* de 0,333 indicando falha na geração. Nesse caso, o LLM respondeu corretamente (4 anos), mas foi penalizado pelo Ragas por acrescentar um comentário dizendo que não encontrou um tempo máximo para a conclusão do curso.

A *Faithfulness* foi 1,000 em 22 questões, entretanto na Q17 (prazo de empréstimo) a resposta do LLM informou não ter encontrado o prazo e recebeu 1,000 em *Faithfulness* e 0 nas demais métricas. Isso ocorre pois o mecanismo RAG não recuperou a tabela que continha a resposta, apesar desse dado estar presente na base. Esse caso também ilustra que a fundamentação em documentos pode coexistir com a ausência da informação solicitada, causada nesse contexto pelo desempenho do sistema.

A *Answer Accuracy* foi igual ou inferior a 0,500 em 24 questões e atingiu 1,000 em apenas cinco. Em 17 das 24 a cobertura foi 1,000, e em 12 a *Faithfulness* foi no mínimo 0,900, respostas fundamentadas em contextos completos, mas divergentes do gabarito. A baixa pontuação em Answer Accuracy inclui múltiplos aspectos como perguntas não respondidas e dados omitidos por parte do LLM de resposta em 11 questões. Porém na maior parte da amostra (23 questões), o fator mais influente foi o rigor do LLM juíz do Ragas, pois o gabarito foi construído com respostas sintetizadas. Por outro lado, o *chatbot* gerou não só a resposta esperada como também incluiu informações adjacentes pautadas nos trechos recuperados. Mesmo sendo incrementos alinhados ao contexto e assertivos, foram penalizados pelo Ragas por não estarem no gabarito.

5. Considerações finais

Vale ressaltar que o estudo apresenta limitações decorrentes do número de perguntas e do uso de outro modelo de linguagem como avaliador. Foi selecionada uma resposta por questão, sem replicações para estimar variância. Portanto, os resultados caracterizam o conjunto observado, sem considerar estabilidade entre execuções ou cobertura de consultas ambíguas e conversas prolongadas. A aplicação do Ragas permitiu examinar separadamente a recuperação documental e a geração de respostas do *chatbot*. O principal resultado foi evidenciar que respostas fundamentadas nos documentos podem não fornecer a informação esperada pelo usuário, e que a maioria das notas baixas de *Answer Accuracy* ocorreu com recuperação completa, sugerindo priorizar a geração e a ordenação dos contextos. A seleção das questões foi por conveniência, refletindo a experiência empírica do autor.

Referências

Es, S., James, J., Espinosa Anke, L., and Schockaert, S. (2024). RAGAs: Automated evaluation of retrieval augmented generation. In *Proceedings of EACL: System Demonstrations*, pages 150–158. Association for Computational Linguistics.

Ragas (2023). List of available metrics. `https://docs.ragas.io/en/v0.4.3/concepts/metrics/available_metrics/`. Acesso em: 21 set. 2026.