A Comparative Analysis of LLMs in a FAQ Chatbot: A Case Study

Samara Monteiro Xavier[1], **Cesar Yuri e Silva Santos**[1],
Anderson Pinheiro Cavalcanti[12], **Péricles Miranda**[1], **Rafael Ferreira Mello**[12]

[1]Universidade Federal Rural de Pernambuco (UFRPE)
[2]CESAR School

{samara.xavier,cesar.yuri,anderson.pinheiro}@ufrpe.br
{pericles.miranda,rafael.mello}@ufrpe.br

Abstract. *This paper presents an experimental study evaluating Large Language Models (LLMs) in a FAQ chatbot. The system adopts a Retrieval-Augmented Generation (RAG) approach and considers different user profiles. The evaluation combines lexical, semantic, and LLM-as-a-judge metrics, as well as operational indicators such as latency and cost. Results show that while lexical metrics penalize verbose responses, semantic and judge-based metrics better capture response quality. The study highlights trade-offs between performance, cost, and adaptability to user profiles.*

Resumo. *Este artigo apresenta um estudo experimental sobre o uso de modelos de linguagem de grande porte (LLMs) em um chatbot de FAQ. O sistema utiliza uma abordagem de geração aumentada por recuperação (RAG) e considera diferentes perfis de usuário. A avaliação combina métricas léxicas, semânticas e baseadas em LLM como juiz, além de indicadores operacionais como latência e custo. Os resultados mostram que métricas léxicas penalizam respostas mais detalhadas, enquanto métricas semânticas e baseadas em julgamento capturam melhor a qualidade das respostas. O estudo evidencia trade-offs entre desempenho, custo e adaptação aos perfis de usuário.*

1. Introdução

Os avanços recentes em Large Language Models (LLMs) têm transformado significativamente a forma como sistemas conversacionais são desenvolvidos e utilizados em diferentes domínios de aplicação, como educação, pesquisa e saúde. Esses modelos apresentam elevada capacidade de compreender, gerar e manipular linguagem natural, sendo amplamente utilizados na construção de chatbots e assistentes virtuais [Singh and Namin 2025].

Chatbots baseados em LLMs apresentam vantagens importantes em relação a abordagens tradicionais baseadas em regras ou recuperação de informação, principalmente por sua capacidade de compreender linguagem natural, lidar com variações linguísticas e gerar respostas mais naturais e contextualizadas. Além disso, tais sistemas contribuem para a melhoria da experiência do usuário e para a oferta de serviços mais ágeis e personalizados [Soares and Silva 2025]. Nesse contexto, a coexistência entre arquiteturas baseadas em regras, recuperação, geração e abordagens híbridas exige uma reavaliação dos métodos de controle de qualidade, pois cada paradigma oferece trade-offs distintos de fluidez, precisão factual, segurança e adaptabilidade

[Papadopoulos et al. 2025]. Em aplicações reais, como sistemas de atendimento ao cliente, ainda são observados problemas relacionados à precisão, adequação contextual e confiabilidade das respostas geradas [Gomes et al. 2025].

Um dos principais desafios associados ao uso de LLMs em sistemas conversacionais é a análise da qualidade das respostas geradas. Diferentemente de tarefas tradicionais de processamento de linguagem natural, nas quais existe uma resposta única ou claramente definida, tarefas de geração textual frequentemente admitem múltiplas respostas semanticamente válidas para uma mesma pergunta, o que torna o processo de avaliação mais complexo [Xu et al. 2023]. Nesse contexto, estudos apontam a ausência de um framework universal para avaliação, destacando a necessidade de múltiplas métricas e abordagens complementares [Singh and Namin 2025].

Diante desses desafios, diversas abordagens de avaliação têm sido propostas na literatura, destacando-se métricas automáticas clássicas, como BLEU e ROUGE, amplamente utilizadas para medir similaridade entre textos gerados e respostas de referência [Papineni et al. 2002] [Lin 2004], embora apresentem limitações conhecidas em tarefas conversacionais [Liu et al. 2016]. Mais recentemente, métricas baseadas em representações semânticas profundas, como BERTScore, têm sido também empregadas para capturar similaridades semânticas mais complexas entre textos, superando limitações das métricas puramente lexicais [Zhang et al. 2019]. Além disso, abordagens recentes têm explorado o uso de modelos de linguagem como avaliadores automáticos (LLM-as-a-judge), buscando aproximar a análise automática dos julgamentos humanos e aumentar a escalabilidade do processo avaliativo [Liu et al. 2023].

Apesar desses avanços metodológicos, métricas automáticas podem apresentar limitações em sua correlação com julgamentos humanos em textos longos, sendo necessário adotar abordagens de avaliação multifacetadas que considerem aspectos como coerência, factualidade e completude [Xu et al. 2023]. No caso específico de avaliadores baseados em LLMs, estudos recentes mostram ainda que o processo de julgamento automático pode incorporar vieses próprios do avaliador, que afetam a confiabilidade da avaliação e podem, inclusive, ser explorados para induzir julgamentos incorretos [Chen et al. 2024]. O desempenho de modelos de linguagem pode variar significativamente dependendo do contexto de aplicação, tipo de tarefa e perfil do usuário atendido pelo sistema conversacional, reforçando a necessidade de avaliações específicas para cada cenário [Singh and Namin 2025]. Com base nesse contexto, este artigo apresenta um estudo experimental sobre o uso de diferentes LLMs em um chatbot de FAQ baseado em RAG, condicionado por quatro perfis de usuário. As principais contribuições do trabalho são: (i) a definição de um cenário experimental realista para avaliação de respostas em FAQ com perfis diferentes; (ii) a comparação sistemática entre modelos comerciais e open-source sob um mesmo pipeline de recuperação e geração; (iii) a adoção de uma avaliação multimétrica que combina métricas lexicais, semânticas, LLM-as-a-judge e indicadores operacionais; e (iv) a análise dos trade-offs entre qualidade das respostas, adequação ao perfil e custo de uso em APIs. Para atingir os objetivos propostos e orientar a condução metodológica deste estudo, foram definidas as seguintes questões de pesquisa (RQs - Research Questions):

RQ1. Como diferentes modelos de linguagem de larga escala (LLMs) se comportam na tarefa de respostas a perguntas em um chatbot baseado em FAQ, considerando a

aderência ao perfil do usuário e a qualidade textual das respostas?

RQ2. Quais são os trade-offs entre qualidade das respostas, latência e custo de uso entre os modelos avaliados?

2. Trabalhos Relacionados

A literatura recente tem explorado intensamente a transição de sistemas conversacionais tradicionais para arquiteturas baseadas em modelos generativos de larga escala. [Singh and Namin 2025] apresentam uma revisão ampla sobre chatbots integrados a LLMs, com ênfase em teste, avaliação e desempenho, enquanto [Luo et al. 2022] analisam o estado da arte no desenvolvimento de chatbots, evidenciando lacunas relacionadas à integração entre técnicas computacionais e contextos de aplicação. Essa mudança de paradigma é corroborada por [Foosherian et al. 2023], que investigam como a integração de LLMs pode expandir as capacidades de agentes baseados em pipeline, mitigando quebras conversacionais e aumentando a capacidade de lidar com solicitações complexas e fora do escopo original. No contexto brasileiro, essa transição também se reflete: enquanto [Gomes et al. 2018] propuseram um chatbot de FAQ para estudantes baseado em classificação de intenção, abordagens mais recentes adotam geração aumentada por recuperação [Kuratomi et al. 2024].

No domínio específico de ambientes de FAQ (Frequently Asked Questions), estudos têm focado na precisão técnica e na confiabilidade da informação. [Nusair et al. 2025] avaliaram o desempenho de diferentes LLMs ao responderem perguntas frequentes sobre cirurgia de catarata, utilizando uma escala de cinco pontos (1 a 5) para medir precisão, completude e clareza. Os resultados indicaram um desempenho satisfatório geral, embora modelos como LLAMA 4 e Copilot tenham apresentado pontuações médias inferiores em comparação aos demais competidores. De forma semelhante, [Fan et al. 2025] analisaram a acurácia de chatbots em FAQs sobre câncer do colo do útero, ressaltando que, apesar das respostas satisfatórias, uma atitude cautelosa e a validação por especialistas humanos continuam sendo essenciais, especialmente em questões relacionadas a protocolos de tratamento.

Além da precisão técnica, outra linha relevante de pesquisa diz respeito à personalização e a adaptação das respostas ao perfil do usuário, que representam uma fronteira crescente na área. [Jiang et al. 2025] introduziram o benchmark PERSONAMEM, focado em avaliar a capacidade dos LLMs de realizar o perfilamento dinâmico de usuários e gerar respostas personalizadas em larga escala. O estudo evidencia que os modelos atuais ainda enfrentam dificuldades para rastrear a evolução dinâmica dos perfis de usuários ao longo do tempo. Em paralelo, [Zhao et al. 2025], com o benchmark PersonaLens, mostram que a capacidade de personalização varia consideravelmente entre assistentes conversacionais e permanece um desafio importante em cenários orientados a tarefas.

A infraestrutura para o desenvolvimento desses sistemas também tem sido simplificada para usuários técnicos e não técnicos. [Salim et al. 2025] propuseram um construtor de chatbots de FAQ baseado em IA generativa que utiliza técnicas de Geração Aumentada por Recuperação (RAG) e ajuste fino (fine-tuning) para garantir que a recuperação de informações organizacionais seja precisa e contextualizada. Para ambientes acadêmicos, [Attigeri et al. 2024] realizaram uma análise comparativa de modelos de Processamento

de Linguagem Natural (PLN) voltados para chatbots de universidades, concluindo que tais ferramentas são eficazes para centralizar informações oficiais e reduzir discrepâncias comumente encontradas em fontes dispersas como sites e folhetos.

O uso de abordagens como LLM-as-a-judge tem sido adotado como alternativa para escalar esse processo, como demonstrado por [Liu et al. 2023], embora estudos mais recentes apontem que tanto humanos quanto LLMs apresentam vieses sistemáticos, comprometendo a confiabilidade das avaliações [Chen et al. 2024]. No que se refere à avaliação automática, BLEU e ROUGE permanecem como baselines úteis [Papineni et al. 2002] [Lin 2004], mas sua correlação com julgamentos humanos é bem limitada em tarefas conversacionais e de geração aberta [Liu et al. 2016] [Xu et al. 2023].

Em síntese, a literatura recente evidencia avanços importantes no uso de LLMs para sistemas conversacionais e ambientes de FAQ, bem como no desenvolvimento de métodos mais abrangentes de avaliação. No entanto, ainda são escassos os estudos que analisam, de forma integrada, modelos distintos em um chatbot de FAQ condicionado por perfis de usuário, considerando simultaneamente qualidade textual, adequação comunicativa, segurança e custo operacional. É nesse espaço que o presente trabalho se insere.

3. Metodologia

3.1. Descrição do Sistema

O sistema avaliado consiste em um chatbot de FAQ, desenvolvido em Python, utilizando arquitetura baseada em Flask. O chatbot segue um modelo de recuperação e geração de respostas (Retrieval-Augmented Generation – RAG), no qual as respostas são geradas com base em uma base de conhecimento composta por documentos em formato PDF. A base de conhecimento é processada por meio de extração textual utilizando a biblioteca PyPDF2, seguida de segmentação em trechos (chunks) com tamanho de 1000 caracteres e sobreposição (overlap) de 200 caracteres. Essa configuração corresponde a um padrão amplamente adotado em sistemas RAG, equivalendo a uma sobreposição de 20%, valor situado na faixa recomendada de 10–20% para preservação de contexto entre trechos e foi empregada de forma semelhante por [Narimissa and Raithel 2024], que demonstram a adequação dessa estratégia para manter a integridade contextual na recuperação. Esses blocos foram indexados localmente em um banco de dados vetorial FAISS, o qual é reconstruído automaticamente caso haja alterações nos arquivos de contexto.

A recuperação de informações utiliza um mecanismo de busca baseado em similaridade vetorial, retornando os k = 5 trechos mais relevantes por consulta. Esses trechos são utilizados como contexto de entrada para o modelo de linguagem por meio de uma cadeia conversacional (*ConversationalRetrievalChain*), com memória de curto prazo (window memory) limitada às últimas interações do usuário. O sistema incorpora ainda um mecanismo de controle de fluxo baseado em estado, no qual o usuário seleciona um perfil específico, influenciando diretamente o escopo, o tom e o nível de detalhamento das respostas geradas. Esse mecanismo é operacionalizado por meio de restrições explícitas no prompt, garantindo que as respostas sejam geradas exclusivamente com base no contexto permitido. Por questões de privacidade e anonimização, os perfis são referenciados neste trabalho como Página Geral, Perfil 1, Perfil 2 e Perfil 3. Essa anonimização não impacta a estrutura experimental nem a interpretação dos resultados, uma vez que os perfis mantêm suas características funcionais no sistema.

Os quatro perfis representam níveis distintos de acesso e especialização informacional. A Página Geral reúne conteúdos públicos; o Perfil 1 adiciona informações específicas para usuários autenticados; o Perfil 3 concentra orientações operacionais mais especializadas; e o Perfil 2 corresponde ao nível mais abrangente, incorporando os conteúdos dos demais perfis. Em termos experimentais, essa organização implica que o conjunto de documentos recuperáveis varia conforme o perfil selecionado, alterando o escopo disponível para geração da resposta. A Figura 1 mostra o diagrama da arquitetura proposta no chatbot.

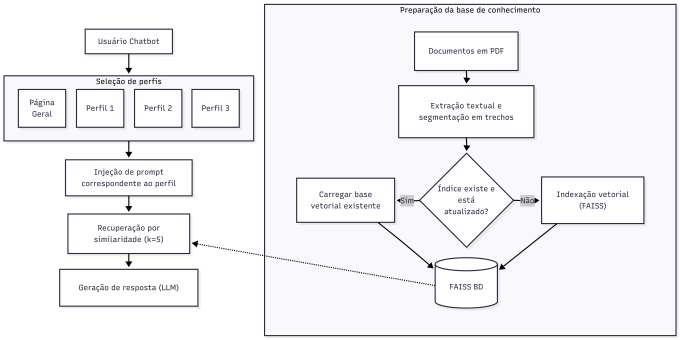


Figura 1. Arquitetura do sistema proposto.

3.2. Bases do Experimento

A variável independente do experimento é o modelo de linguagem utilizado na geração das respostas. Foram avaliados diferentes modelos comerciais e open-source, incluindo variações de modelos da família GPT, Gemini, Llama e Gemma: gemini-2.5-flash; GPT-5 Mini; GPT-5.2; GPT-4o; Llama 3.2:3b; e Gemma 3:4b. A execução experimental para comparação dos geradores manteve o modelo de representação vetorial fixado em Gemini Embeddings (gemini-embedding-001). Dessa forma, garantiu-se que as variações no texto gerado fossem estritamente derivadas do modelo LLM utilizado.

Os modelos open-source foram executados localmente em infraestrutura computacional equipada com processador AMD Ryzen 5 5500, placa-mãe MSI A520M-A PRO, GPU AMD Radeon RX 580 com 8 GB de memória dedicada e 16 GB de memória RAM, utilizando o framework Ollama para inferência local. Todos os modelos foram executados com temperatura configurada em 0, buscando reduzir variabilidade estocástica e aumentar a reprodutibilidade dos experimentos.

O conjunto experimental foi construído a partir de uma base de conhecimento composta por 39 perguntas frequentes (FAQ), distribuídas entre quatro perfis de usuário: Página Geral, Perfil 1, Perfil 2 e Perfil 3. A distribuição ocorreu como mostrado na Tabela 1. Por questões de privacidade, os nomes originais dos perfis não são apresentados neste artigo. Ainda assim, a distribuição das FAQs por perfil foi mantida na caracterização do experimento, uma vez que essa organização influencia diretamente o escopo da recuperação de informação e a complexidade das respostas geradas.

A estrutura do experimento reflete uma lógica hierárquica de acesso, na qual perfis mais abrangentes incorporam conteúdos acessíveis a perfis anteriores. Como consequência, a interseção entre perguntas de diferentes perfis gera casos adicionais de teste, ampliando o número total de avaliações realizadas. Essa característica é particularmente

Tabela 1. Distribuição dos casos de teste por perfil

Perfil	Nº de casos	Percentual	Tam. médio da resposta de referência
Página Geral	7	17.9%	30.86
Perfil 1	11	28.2%	19.45
Perfil 2	13	33.3%	15.08
Perfil 3	8	20.5%	13.25
Total	**39**	**100%**	**18.77**

relevante, pois introduz variações no nível de complexidade das consultas e no contexto disponível para geração das respostas.

As respostas de referência (*ground truth*) caracterizam-se por serem diretas e concisas, apresentando um tamanho médio geral de 18.7 palavras (120.4 caracteres) por resposta. A granularidade das informações variou conforme o perfil: o acesso público (Página Geral) exigiu instruções mais extensas para orientação do cidadão (média de 30.8 palavras), enquanto as dúvidas operacionais restritas aos perfis técnicos demandaram orientações mais breves e pontuais (médias entre 13 e 20 palavras).

3.3. Métricas de Avaliação

A avaliação das respostas combinou métricas lexicais, semânticas, baseadas em julgamento automático e indicadores operacionais. Foram utilizados BLEU e ROUGE-L para medir proximidade lexical e estrutural entre as respostas geradas e as respostas de referência, BERTScore F1 para capturar similaridade semântica e GEval, implementado com a biblioteca DeepEval. A avaliação com GEval foi conduzida por dois juízes independentes, Gemini-2.5-flash e GPT-5-mini, ambos com temperatura zero, a fim de verificar a robustez das pontuações frente ao avaliador escolhido. No presente estudo, a aderência ao perfil foi operacionalizada no GEval por meio de três critérios: (i) respeito ao escopo informacional do perfil selecionado; (ii) adequação do nível de detalhamento e do tom da resposta ao público-alvo; e (iii) ausência de informações exclusivas de outros perfis. Além disso, foram medidos o tempo médio de resposta e o custo estimado de API, de modo a analisar o trade-off entre qualidade, latência e viabilidade de implantação.

3.4. Procedimento Experimental

Cada caso de teste foi executado sob as mesmas condições de recuperação e geração, variando apenas o modelo de linguagem utilizado. Para cada pergunta, o pipeline completo foi acionado: seleção do perfil correspondente, recuperação dos cinco trechos mais similares, geração da resposta pelo LLM e armazenamento dos resultados em formato CSV. Os arquivos gerados registraram, para cada execução, a pergunta, a resposta de referência, a resposta produzida, o modelo utilizado e o perfil associado. Em seguida, as respostas foram submetidas a um pipeline automatizado de avaliação, responsável por calcular as métricas por item e as médias agregadas por modelo e por perfil. Também foram realizadas análises complementares de comprimento textual, a fim de apoiar a interpretação das diferenças entre métricas lexicais, semânticas e baseadas em julgamento automático.

4. Resultados

Nesta seção, apresentamos os resultados experimentais obtidos na avaliação dos modelos de linguagem na tarefa de geração de respostas para o chatbot de FAQ. A análise foi

estruturada para responder diretamente às duas questões de pesquisa do estudo: (i) qualidade das respostas e adequação aos perfis de usuário e (ii) trade-offs entre qualidade, latência e custo. A fim de verificar a robustez da avaliação automática, a métrica GEval foi computada com dois juízes independentes de famílias distintas (Gemini2.5-flash e GPT-5-mini), ambos configurados com temperatura zero e submetidos à mesma rubrica. A Tabela 2 resume as métricas agregadas por modelo, reunindo indicadores de qualidade textual, avaliação semântica e julgamento automático sob os dois juízes.

Tabela 2. Métricas gerais por modelo

Modelo	BLEU	ROUGE-L	BERTScore	GEval$_{Gem}$	GEval$_{GPT}$
gemini-2.5-flash	5.65	32.37	63.88	96.41	**90.00**
llama3.2:3b	7.13	31.84	61.89	75.38	65.90
gemma3:4b	**8.12**	30.59	62.71	81.03	77.95
gpt-4o	7.81	34.14	**65.24**	95.38	88.97
gpt-5-mini	5.72	32.61	63.08	97.18	87.18
gpt-5.2	6.92	**38.18**	64.12	**97.69**	89.23

4.1. Desempenho de Geração Textual e Aderência aos Perfis

A avaliação da qualidade textual, medida pelas métricas BLEU, ROUGE e BERTScore, indicou que modelos da família GPT e o Gemini Flash apresentaram respostas mais consistentes. Os resultados agregados demonstram uma discrepância notável entre as métricas de sobreposição lexical (BLEU e ROUGE) e as métricas de similaridade semântica (BERTScore) e LLM-as-a-judge (GEval).

De forma geral, os modelos apresentaram pontuações baixas na métrica BLEU. Por exemplo, o modelo GPT-4o obteve um BLEU médio geral de 7,81, enquanto o gemini-2.5-flash registrou 5,65. Essa baixa sobreposição de n-gramas não indica necessariamente má qualidade, mas sim uma divergência de estilo: a análise das razões de comprimento textual (*ratio words*) revelou que os modelos geraram respostas consideravelmente mais longas do que as referências (o GPT-4o gerou cerca de 4,39 vezes mais palavras, e o gemini-2.5-flash, 5,32 vezes mais). Mesmo fornecendo informações no prompt para o modelo responder diretamente, os LLMs tendem a formular respostas mais conversacionais e detalhadas, o que penaliza o BLEU face a respostas de referência tipicamente concisas e diretas. Com relação à métrica ROUGE, o modelo GPT-5.2 obteve o maior desempenho em ROUGE-L médio geral (38,18), enquanto modelos como GPT-4o, GPT-5-mini e Gemini-2.5-flash apresentaram pontuações de ROUGE-L moderadas (entre 32,37 e 34,14). Por outro lado, quando avaliados sob a ótica da captura de significado e utilidade, o desempenho é significativamente superior. O modelo GPT-4o liderou a métrica semântica, atingindo um BERTScore médio geral de 65,24, seguido de perto pelo GPT-5.2 (64,12) e gemini-2.5-flash (63,88). Isso sugere uma boa fidelidade aos conceitos presentes na base de conhecimento.

No que diz respeito à adequação das respostas aos diferentes perfis (Página Geral, Perfil 1, Perfil 2 e Perfil 3), a métrica GEval foi computada com os dois juízes. Sob o juiz Gemini, os modelos GPT-5.2, GPT-5-mini, Gemini-2.5-flash e GPT-4o apresentaram médias superiores a 95 pontos em uma escala de 0 a 100, com o GPT-5.2 liderando (97,69). O juiz GPT-5-mini, mais rigoroso, deslocou as pontuações para baixo (médias entre 87 e 90 pontos para esses mesmos modelos), mas preservou a separação clara em relação aos modelos *open-source*; nessa condição, o gemini-2.5-flash obteve a

maior média (90,00). A análise qualitativa das justificativas (razões do GEval) aponta que os modelos não só recuperaram a informação correta, como também adaptaram a estrutura da resposta com clareza, adicionando, de forma pertinente, contextos de suporte que enriquecem a experiência do usuário. Em contraste, os modelos Llama 3.2:3b e Gemma 3:4b obtiveram pontuações consistentemente menores sob ambos os juízes (Gemini: 75,38 e 81,03; GPT-5-mini: 65,90 e 77,95, respectivamente), sugerindo maior dificuldade em atender plenamente aos critérios de qualidade complexos definidos na rubrica do juiz.

Quanto à aderência aos perfis específicos, observou-se, sob o juiz Gemini, uma tendência de queda de desempenho nos perfis mais especializados: na maioria dos modelos, a "Página Geral" superou o "Perfil 3", indicando que a especialização da resposta pelo perfil introduz uma camada adicional de complexidade para os geradores. Contudo, essa tendência mostrou-se dependente do juiz: sob o GPT-5-mini, o padrão é menos nítido, e diversos modelos alcançam suas maiores pontuações em perfis intermediários (Perfil 1) ou mesmo no Perfil 3. Esse contraste sugere que parte da penalização atribuída à especialização por perfil pode refletir preferências do avaliador, reforçando a utilidade de múltiplos juízes e a importância de métricas baseadas em julgamento semântico e qualitativo para avaliação em cenários personalizados. A Tabela 3 apresenta a variação das métricas por perfil de usuário, incluindo o GEval médio entre os dois juízes.

A comparação entre os dois juízes sustenta a confiabilidade da avaliação automática. Primeiro, há forte concordância quanto à ordenação dos modelos ($r \approx 0{,}96$ no nível de modelo), com separação consistente entre os modelos comerciais e os modelos *open-source* de menor porte, que ocuparam as duas últimas posições sob ambos os avaliadores. Segundo, não se observou indício de favorecimento de nenhum juiz a modelos de sua própria família: o gemini-2.5-flash, longe de ser inflado pelo juiz Gemini, alcançou sua melhor colocação relativa sob o juiz GPT-5-mini (90,00, a maior pontuação nessa condição), e o juiz GPT-5-mini atribuiu ao gerador gpt-5-mini a maior redução de pontuação entre todos os modelos (de 97,18 para 87,18). Terceiro, o juiz GPT-5-mini mostrou-se sistematicamente mais rigoroso, com queda média em torno de 7 pontos, o que reflete diferenças de calibração entre avaliadores sem alterar a estrutura relativa dos resultados. Ressalva-se, todavia, que a concordância é mais robusta na distinção entre modelos comerciais e *open-source* do que na ordenação fina entre os modelos comerciais de topo: enquanto o juiz Gemini posiciona o GPT-5.2 como melhor modelo, o juiz GPT-5-mini atribui a melhor média ao gemini-2.5-flash.

De modo geral, o BERTScore mostrou-se útil para medir a fidelidade semântica à resposta de referência, enquanto o GEval captura de forma mais adequada a competência comunicativa dos modelos, especialmente em cenários nos quais a resposta correta pode ser expressa com maior liberdade linguística. Esses resultados indicam que, embora os modelos apresentem boa capacidade de adaptação geral, a especialização por perfil introduz desafios adicionais que não são plenamente capturados por métricas lexicais.

4.2. Análise de Custos e Eficiência Operacional

A viabilidade de implantação de um sistema conversacional em ambientes de produção exige um equilíbrio rigoroso entre a qualidade das respostas e o custo computacional associado. Para responder à RQ2, avaliamos o tempo de resposta (latência média) e o custo com API estimado (por mil tokens) para cada modelo. A relação entre qualidade e

Tabela 3. BERTScore F1 e GEval por perfil (juízes Gemini e GPT-5-mini)

Modelo	Perfil	BERTScore F1	GEval_{Gem}	GEval_{GPT}	GEval médio
gemini-2.5-flash	Página Geral	**72.75**	98.57	90.00	**94.29**
	Perfil 1	64.46	97.27	88.18	92.73
	Perfil 2	61.79	96.92	89.23	93.08
	Perfil 3	58.72	92.50	93.75	93.13
llama3.2:3b	Página Geral	**65.80**	67.14	62.86	65.00
	Perfil 1	62.66	82.73	72.73	**77.73**
	Perfil 2	60.14	75.38	62.31	68.85
	Perfil 3	60.25	72.50	65.00	68.75
gemma3:4b	Página Geral	**69.41**	80.00	74.29	77.15
	Perfil 1	63.23	86.36	86.36	**86.36**
	Perfil 2	62.33	82.31	76.15	79.23
	Perfil 3	56.76	72.50	72.50	72.50
gpt-4o	Página Geral	**73.24**	97.14	91.43	**94.29**
	Perfil 1	66.02	98.18	89.09	93.64
	Perfil 2	63.16	95.38	88.46	91.92
	Perfil 3	60.53	90.00	87.50	88.75
gpt-5-mini	Página Geral	**71.22**	97.14	74.29	85.72
	Perfil 1	63.19	99.09	94.55	**96.82**
	Perfil 2	61.38	98.46	85.38	91.92
	Perfil 3	58.59	92.50	91.25	91.88
gpt-5.2	Página Geral	**73.98**	98.57	85.71	92.14
	Perfil 1	63.97	99.09	91.82	95.46
	Perfil 2	62.34	99.23	92.31	**95.77**
	Perfil 3	58.62	92.50	83.75	88.13

custo dos modelos pode ser observada na Figura 2, que apresenta um gráfico de dispersão entre o GEval médio e o custo por caso. Percebe-se a existência de um trade-off entre custo e qualidade, no qual modelos como GPT-5.2 e GPT-4o apresentam alta qualidade a custos moderados, enquanto alternativas *open-source* oferecem custo de API reduzido com impacto significativo na qualidade.

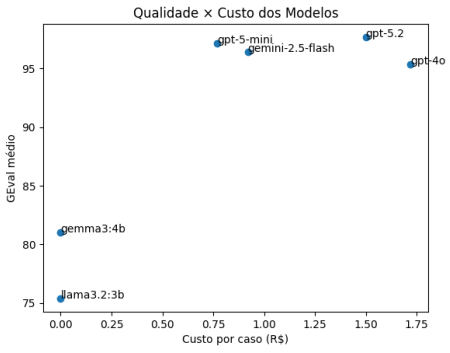


Figura 2. Relação entre qualidade e custo por caso dos modelos avaliados

No plano operacional, observou-se um contraste importante entre latência e custo entre os modelos avaliados. Entre os modelos com API paga, o gpt-5.2 apresentou o menor tempo médio de resposta (2,65 s), seguido por gpt-4o (3,39 s), enquanto gpt-5-mini (17,80 s) e Gemini-2.5-flash (36,47 s) exibiram maior latência. Em termos de custo estimado com API, considerando os volumes de tokens observados nos 39 casos de cada

modelo e os preços informados, o gpt-5-mini foi o modelo pago de menor gasto, com custo total aproximado de R$ 0,77 no conjunto avaliado (cerca de R$ 0,0155 por caso), superando em economia o gpt-5.2 (R$ 1,50), o gpt-4o (R$ 1,72) e o Gemini 2.5 Flash, cujo custo mínimo estimado foi de R$ 0,92. Esses resultados indicam que, sob a perspectiva de eficiência operacional, o gpt-5.2 se destacou em velocidade, enquanto o gpt-5-mini ofereceu a melhor relação de economia entre os modelos pagos.

5. Conclusão

Este artigo apresentou um estudo experimental sobre o uso de diferentes LLMs em um chatbot de FAQ baseado em RAG, condicionado por perfis de usuário. Os resultados reforçam que a avaliação nesse cenário não pode depender apenas de métricas lexicais: BLEU e ROUGE-L apresentaram valores modestos, enquanto BERTScore e GEval capturaram melhor a adequação semântica e comunicativa dos modelos comerciais, reflexo do estilo mais explicativo dos LLMs frente às respostas curtas da base de FAQ. A adaptação ao perfil mostrou-se um desafio adicional, ainda que menos uniforme do que sugeriria um único avaliador: sob o juiz Gemini, a Página Geral tendeu a superar o Perfil 3, padrão que se dissolveu sob o GPT-5-mini, indicando que parte da penalização por especialização pode refletir preferências do avaliador.

A adoção de dois juízes independentes para o GEval qualifica essas conclusões. A forte concordância de ordenação e a ausência de indícios de favorecimento de cada juiz a modelos de sua própria família aumentam a confiança nas pontuações e atenuam a preocupação com *self-preference bias*; em contrapartida, o melhor modelo isolado variou conforme o avaliador (GPT-5.2 sob o Gemini, gemini-2.5-flash sob o GPT-5-mini), ao passo que a separação entre modelos comerciais e *open-source* permaneceu estável, constituindo o achado mais sólido do estudo. No plano operacional, não houve modelo superior em todas as dimensões: o GPT-5.2 ofereceu o melhor equilíbrio entre qualidade e latência e o GPT-5-mini o menor custo entre os pagos, sugerindo que a escolha para implantação real deve ponderar simultaneamente qualidade, tempo de atendimento e orçamento. Como trabalhos futuros, pretende-se ampliar o conjunto de casos, incorporar avaliação humana como referência para calibrar os juízes automáticos e investigar métricas mais diretamente relacionadas à fidelidade ao contexto recuperado.

6. Limitações

Este estudo apresenta limitações que devem ser consideradas e que indicam direções para trabalhos futuros. Primeiro, o experimento foi conduzido com um conjunto restrito de casos de teste, derivados de um número limitado de perguntas frequentes de um domínio específico; embora adequado a um estudo exploratório, esse volume limita a generalização estatística dos resultados para domínios de maior complexidade. Segundo, a avaliação qualitativa baseou-se em modelos de linguagem como juízes (LLM-as-a-judge). Embora a recomputação com um segundo juiz de família distinta não tenha revelado indícios de autopreferência e tenha apresentado forte concordância de ordenação, ambos os avaliadores podem compartilhar vieses sistemáticos de estilo, verbosidade ou formatação [Chen et al. 2024]; ademais, o juiz primário (Gemini) pertence à mesma família de um dos modelos avaliados e do modelo de *embeddings* utilizado na recuperação. Dessa forma, os resultados de GEval devem ser interpretados como complementares às métricas

léxicas e semânticas, e não como substitutos da avaliação humana. Terceiro, o pipeline de recuperação foi mantido fixo, incluindo o modelo de *embeddings*, a estratégia de segmentação (*chunking*) e o banco vetorial, com o objetivo de isolar a variável de geração. Embora metodologicamente justificável, essa decisão pode não refletir o desempenho ótimo que cada LLM atingiria caso o mecanismo de recuperação fosse ajustado às suas características [Papadopoulos et al. 2025, Foosherian et al. 2023]. Por fim, os valores de latência e custo representam um recorte do momento de execução dos experimentos e estão sujeitos a variações decorrentes de carga nos servidores, latência de rede e atualizações nas políticas de precificação.

Disponibilidade de artefatos

Os artefatos deste estudo estão disponíveis no link: `https://anonymous.4open.science/r/faq-chatbot-eniac-84AD/`.

Referências

Attigeri, G., Agrawal, A., and Kolekar, S. V. (2024). Advanced nlp models for technical university information chatbots: development and comparative analysis. *IEEE Access*, 12:29633–29647.

Chen, G. H., Chen, S., Liu, Z., Jiang, F., and Wang, B. (2024). Humans or llms as the judge? a study on judgement bias. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8301–8327.

Fan, J., Xiao, W., Yan, Z., and Ouyang, Q. (2025). Accuracy of ai chatbots in answering frequently asked questions on cervical cancer. *Frontiers in Artificial Intelligence*, 8:1655303.

Foosherian, M., Purwins, H., Rathnayake, P., Alam, T., Teimao, R., and Thoben, K.-D. (2023). Enhancing pipeline-based conversational agents with large language models. In *Proceedings of the 1st workshop on taming large language models: controllability in the era of interactive assistants!*, pages 56–67.

Gomes, J. a. E. A., Prudêncio, R. B. C., and Nascimento, A. C. A. (2018). Chatbot to support frequently asked questions from students in higher education institutions. In *Anais do XV Encontro Nacional de Inteligência Artificial e Computacional (ENIAC)*, Porto Alegre, RS, Brasil. SBC.

Gomes, T., Regino, A. G., Caus, R., Sotelo, V., and dos Reis, J. C. (2025). Multi-agent llm approach for moderating e-commerce customer service responses. In *Brazilian Symposium on Multimedia and the Web (WebMedia)*, pages 349–357. SBC.

Jiang, B., Hao, Z., Cho, Y.-M., Li, B., Yuan, Y., Chen, S., Ungar, L., Taylor, C. J., and Roth, D. (2025). Know me, respond to me: Benchmarking llms for dynamic user profiling and personalized responses at scale. *arXiv preprint arXiv:2504.14225*.

Kuratomi, G., Pirozelli, P., Cozman, F. G., and Peres, S. M. (2024). A RAG-Based institutional assistant. In *Anais do XXI Encontro Nacional de Inteligência Artificial e Computacional (ENIAC)*, Porto Alegre, RS, Brasil. SBC.

Lin, C.-Y. (2004). Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Liu, C.-W., Lowe, R., Serban, I. V., Noseworthy, M., Charlin, L., and Pineau, J. (2016). How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 2122–2132.

Liu, Y., Iter, D., Xu, Y., Wang, S., Xu, R., and Zhu, C. (2023). G-eval: Nlg evaluation using gpt-4 with better human alignment, 2023. *arXiv preprint arXiv:2303.16634*, 12:1.

Luo, B., Lau, R. Y., Li, C., and Si, Y.-W. (2022). A critical review of state-of-the-art chatbot designs and applications. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 12(1):e1434.

Narimissa, E. and Raithel, A. (2024). Exploring information retrieval landscapes: An investigation of a novel evaluation techniques and comparative document splitting methods. *arXiv preprint arXiv:2409.08479*.

Nusair, O., Asadigandomani, H., Osuorah, I., Sanchez, J., and Soleimani, M. (2025). Evaluating the reliability of large language models in answering faqs for cataract surgery. *Digital Health*, 11:20552076251406304.

Papadopoulos, T., Alexopoulos, C., and Charalabidis, Y. (2025). Evaluating chatbot architectures for public service delivery: balancing functionality, safety, ethics, and adaptability. *Frontiers in Political Science*, 7:1601440.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Salim, M. S., Hossain, S. I., Jalal, T., Bose, D. K., and Basher, M. J. I. (2025). Llm based qa chatbot builder: A generative ai-based chatbot builder for question answering. *SoftwareX*, 29:102029.

Singh, S. U. and Namin, A. S. (2025). A survey on chatbots and large language models: Testing and evaluation techniques. *Natural Language Processing Journal*, 10:100128.

Soares, J. R. and Silva, P. N. (2025). Concepção de chatbots: perspectivas e interseções na ciência da informação. *RDBCI: Revista Digital de Biblioteconomia e Ciência da Informação*, 23:e025031.

Xu, F., Song, Y., Iyyer, M., and Choi, E. (2023). A critical evaluation of evaluations for long-form question answering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3225–3245.

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., and Artzi, Y. (2019). Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Zhao, Z., Vania, C., Kayal, S., Khan, N., Cohen, S. B., and Yilmaz, E. (2025). Personalens: A benchmark for personalization evaluation in conversational ai assistants. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 18023–18055.