Lusi: um *chatbot* baseado em Modelos de Linguagem para auxiliar nos atendimentos ao público em departamentos acadêmicos

Alesandro da Silva Santos[1], **Breno Santana Santos**[1],
Raphael Pereira de Oliveira[1], **Marianne Silva**[2]

[1]Departamento de Sistemas de Informação
Universidade Federal de Sergipe – Itabaiana, SE – Brazil

[2]Curso de Sistemas de Informação
Universidade Federal de Alagoas – Penedo, AL – Brazil

{alesandr0,raphael.oliveira}@academico.ufs.br,
breno1005@hotmail.com, marianne.silva@penedo.ufal.br

Resumo. *Departamentos acadêmicos lidam com demandas complexas — a exemplo de dúvidas sobre matrículas e disciplinas —, mas chatbots tradicionais, baseados em regras, falham ao oferecer respostas contextualizadas, gerando frustração e custos operacionais elevados. Desse modo, este estudo propõe o Lusi 2.0, um chatbot baseado em Large Language Models (LLMs) e Retrieval-Augmented Generation (RAG), o qual recupera informações semânticas de Projetos Pedagógicos de Cursos (PPCs) e gera respostas precisas e contextualizadas, reduzindo a necessidade de intervenção manual. Adicionalmente, um experimento controlado comparará sua eficácia com um chatbot baseado em regras, envolvendo, pelo menos, 30 alunos de um departamento universitário. Logo, os resultados permitirão validar sua capacidade de humanizar serviços públicos, substituindo burocracia por diálogos empáticos, além de contribuir para aplicações práticas de LLMs em contextos educacionais.*

1. Introdução

Departamentos acadêmicos de instituições de ensino são pontos centrais para esclarecer dúvidas críticas — de matrículas a requisitos disciplinares —, mas a demanda por atendimento rápido e personalizado muitas vezes esbarra em limitações humanas e tecnológicas. Embora *chatbots* baseados em regras tenham surgido como solução inicial, sua rigidez gerou frustração — respostas genéricas, incapacidade de adaptar-se a contextos e custos elevados de manutenção —, transformando promessas de eficiência em barreiras para seus usuários [Singh et al. 2019].

Nesse contexto, os *Large Language Models* (LLMs) surgem como uma alternativa para tal problema. Capazes de interpretar nuances e aprender com interações, esses modelos permitem desenvolver agentes conversacionais que combinam precisão técnica com uma comunicação mais humana [Raschka 2024, Medeiros et al. 2023a]. Desse modo, este estudo propõe um *chatbot* baseado em LLM para atendimento ao público em departamentos acadêmicos de instituições de ensino. Através de um experimento controlado, espera-se obter evidências de sua eficácia na redução de custos operacionais, na agilização de respostas contextualizadas e na promoção de uma melhor experiência para seus usuários.

2. Trabalhos Relacionados

[Lieb and Goel 2024] desenvolveram o NewtBot, um *chatbot* baseado no GPT-3.5 como tutor personalizado de física para estudantes, com três configurações: genérica ("*baseline*"), contextualizada ("tutor") e específica ("*feedback*"). Em um estudo com 50 alunos alemães, analisaram engajamento e experiência do usuário, identificando que a versão "tutor" obteve as melhores avaliações. Apesar de 72% dos participantes expressarem receios prévios sobre *chatbots* educacionais, 70% relataram disposição para usar a ferramenta. Concluiu-se que adaptações pedagógicas específicas (como o modo "tutor") aumentavam a eficácia e aceitação, indicando que *chatbots* educacionais personalizados poderiam mitigar resistências e apoiar aprendizagem.

[Medeiros et al. 2023a, Medeiros et al. 2023b] avaliaram três abordagens de *chatbots* baseados em modelos de linguagem para resolver consultas em manuais automotivos em PDF: *Doc ChatBot*, *Ask your PDF* e *Question and Answer System*. Por meio de um estudo de caso com os manuais do Ford Fiesta 2015 [Medeiros et al. 2023a] e Ford KA 2008 [Medeiros et al. 2023b], os autores testaram *prompts* (*zero-shot*, *one-shot*, *few-shot*) e parâmetros padronizados (*chunk_size*, *chunk_overlap*) para analisar precisão, custo e experiência do usuário. Os resultados indicaram maior precisão e usabilidade no *Ask your PDF*, economia no *Question and Answer System* (porém, com menor acurácia) e custos elevados no *Doc ChatBot*. Concluiu-se que a escolha recomendada dependia dos critérios prioritários (exatidão, custo ou *interface*), recomendando aprimorar interpretação de elementos visuais e otimização de custos para aplicações práticas.

3. Solução Proposta

Nesta seção, será detalhada a abordagem a ser desenvolvida para auxiliar departamentos acadêmicos no atendimento ao público em geral, o *chatbot* Lusi 2.0. Desse modo, a Figura 1 ilustra o *pipeline* e arquitetura deste agente conversacional.

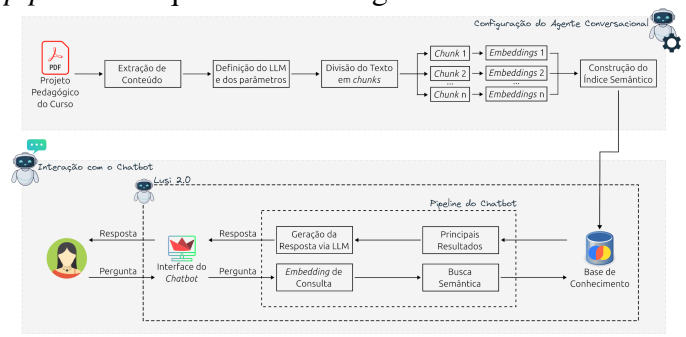

Figura 1. *Pipeline* e Arquitetura do Lusi 2.0

Em termos de arquitetura, utilizou-se a biblioteca Streamlit [Richards 2023] para a construção da *interface* do agente conversacional, a qual, posteriormente, pode ser integrada nas páginas Web de departamentos de qualquer instituição de ensino. Adicionalmente, o banco de dados ChromaDB[1] se responsabiliza pela indexação e recuperação de *embeddings* textuais, possibilitando a realização de buscas semânticas em Projetos Pedagógicos de Cursos (PPC) e o acesso de informações pertinentes. Com relação ao modelo de linguagem utilizado, escolheu-se o Llama 3.2-3B-Instruct[2]. A escolha foi fundamentada nas seguintes justificativas: (i) modelo de estado da arte para *chatbots*; (ii)

[1]https://www.trychroma.com/.

[2]https://huggingface.co/meta-llama/Llama-3.2-3B-Instruct.

comprimento de contexto de 128K *tokens*; e (iii) alta capacidade de geração de texto multilíngue [Alto 2024, Raschka 2024].

Para a construção do agente conversacional, de acordo com a primeira etapa da Figura 1, o processo inicia-se com a escolha e importação em PDF de um PPC de um curso. Em seguida, tem-se a extração do conteúdo desse, bem como a definição dos parâmetros do modelo de linguagem (Llama 3.2-3B-Instruct). Ademais, o texto extraído do PCC é dividido em porções menores, chamadas de *chunks*. Por sua vez, esses fragmentos são convertidos em representações latentes em formato vetorial, denominadas de *embeddings*, as quais capturam o significado semântico dos textos [Raschka 2024]. Por fim, os *embeddings* gerados são persistidos no ChromaDB, gerando um índice semântico e, consequentemente, servindo de base de conhecimento para o agente conversacional. Vale destacar que isso permite que ele recupere informações relevantes de forma efetiva, quando consultado.

Outrossim, na segunda etapa da Figura 1, tem-se o processo de interação de usuários com o *chatbot*. Este inicia-se quando um usuário realiza uma pergunta ao agente. Em seguida, seu texto é, então, convertido em um *embedding* de consulta. Dessa forma, este é utilizado para realizar uma busca por similaridade dentro do índice semântico previamente organizado no ChromaDB. Isto facilita a identificação de correspondências mais relevantes baseadas na semântica. Ainda, a partir desse processo, o agente alicerçado pelo Llama gera uma resposta. Logo, os *chunks* de texto recuperados são classificados por relevância, e os mais alinhados são selecionados para formular a resposta contextualizada e precisa, garantindo uma compreensão eficaz e uma interação natural com o usuário.

No atual momento da pesquisa, a *interface* de conversação está sendo refinada utilizando os princípios de *User Experience* e com base nos *feedbacks* coletados de um teste de usabilidade, previamente realizado com uma pequena amostra aleatória de usuários finais (ver Seção 4.1). Conforme mencionado, o *chatbot* já foi implementado, faltando apenas o refinamento da *interface* de interação. Uma vez finalizada, o próximo passo deste estudo será sua avaliação empírica por meio de um experimento controlado, o qual está detalhado na próxima seção.

4. Metodologia

Esta seção descreve a avaliação experimental da abordagem proposta, a qual segue as diretrizes de [Santos et al. 2018b, Santos et al. 2018a, Wohlin et al. 2024]. Desse modo, o principal objetivo deste estudo é comparar a eficácia de dois agentes conversacionais em responder perguntas específicas de um PPC.

4.1. Planejamento

Com relação à seleção de contexto, o experimento terá como alvo discentes do Departamento de Sistemas de Informação (DSI) — usuários finais —, Campus Itabaiana, Universidade Federal de Sergipe (UFS). Ademais, a questão de pesquisa que precisa ser respondida é "*um chatbot baseado em LLM e RAG, contextualizado com um PPC, é mais eficaz do que um agente baseado em regras pré-definidas em responder a perguntas sobre um determinado PPC?*". Para avaliá-la, serão utilizadas as seguintes métricas: acurácia (% de respostas corretas); precisão (nota de 1 a 5 referente à qualidade da resposta); cobertura (% de perguntas relacionadas ao PPC respondidas corretamente); e grau de satisfação.

Logo, para essa questão de pesquisa, a hipótese a ser confirmada será:

Hipótese nula H_0: o agente conversacional baseado em LLM e RAG possui a mesma precisão e abrangência em respostas a perguntas relacionadas a um PPC, comparado a um *chatbot* baseado em regras pré-definidas.

$$H_0 : performance(chatbot_{LLM}) = performance(chatbot_{regras})$$

Hipótese alternativa H_1: o agente conversacional baseado em LLM e RAG terá precisão e abrangência diferentes em respostas a perguntas relacionadas a um PPC, comparado a um *chatbot* baseado em regras pré-definidas.

$$H_1 : performance(chatbot_{LLM}) \neq performance(chatbot_{regras})$$

Vale ressaltar que H_0 é a hipótese a ser refutada, enquanto H_1 é aquela que se espera corroborar por meio da rejeição de H_0. Ademais, após a definição das hipóteses, inicia-se o processo de seleção de participantes e objetos. Assim, por conveniência, foi determinado que o DSI representaria a população para a realização desta investigação, bem como a amostra, de pelo menos, de 30 indivíduos desse universo seria escolhida ao acaso. Desse modo, será realizado um pedido formal ao DSI, convidando seus alunos a participarem voluntariamente da avaliação experimental.

Adicionalmente, baseado nos *designs* experimentais dos estudos de [Santos et al. 2018b, Santos et al. 2018a], o experimento será projetado num contexto de amostras independentes, em que cada indivíduo avaliará apenas uma das abordagens (*chatbot* baseado em regras pré-definidas ou em LLM e RAG), bem como a atribuição dos participantes aos tratamentos (tipo de agente) será de forma aleatória. Também é importante frisar que os voluntários deverão seguir estritamente um roteiro de interação, onde estarão especificadas perguntas e expectativas de respostas pré-definidas, as quais serão definidas posteriormente com a coordenação do DSI.

Outrossim, quanto à instrumentação, seu processo se dará, inicialmente, com a configuração do ambiente para o experimento, o qual será realizado em um laboratório de informática do DSI, e o planejamento de coleta de dados. Como recursos, utilizar-se-á o tipo de *chatbot* e o roteiro de interação.

4.2. Operação

Esta subseção descreve o processo de execução do experimento. Na etapa de preparação, serão realizados os seguintes passos: (i) ambientação dos participantes, em que estes serão recepcionados, além da apresentação do experimento juntamente com seu principal objetivo. Em seguida, eles preencherão os formulários de caracterização e de consentimento; (ii) sorteio dos indivíduos e de seus tratamentos (tipo de agente conversacional); e (iii) preparação dos computadores do laboratório para a avaliação empírica.

Após a realização dos passos supracitados, iniciar-se-á o experimento de acordo com o *design* planejado. Ademais, o processo de coleta de dados consistirá em duas formas: (i) ao final de uma interação, os participantes responderão um questionário, contido no roteiro de conversação, para avaliar a eficácia de seu agente em resolução de problemas. Ademais, as mensagens trocadas entre os envolvidos (interlocutor-agente) serão armazenadas em uma base de dados para corroborar os acertos e erros dos *chatbots*.

Referências

Alto, V. (2024). *Building LLM Powered Applications: Create intelligent apps and agents with large language models*. Packt Publishing Ltd.

Lieb, A. and Goel, T. (2024). Student interaction with newtbot: An llm-as-tutor chatbot for secondary physics education. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, New York, NY, USA. Association for Computing Machinery.

Medeiros, T., Medeiros, M., Azevedo, M., Silva, M., Silva, I., and Costa, D. G. (2023a). Analysis of language-model-powered chatbots for query resolution in pdf-based automotive manuals. *Vehicles*, 5(4):1384–1399.

Medeiros, T., Medeiros, M., Azevedo, M., Wanderley, E., Silva, M., and Silva, I. (2023b). Análise de chatbots habilitados por modelos de linguagem para resolução de consultas em manuais automotivos no formato pdf. In *XVI Simpósio Brasileiro de Automação Inteligente (SBAI 2023)*, Manaus, Amazonas, Brasil. Sociedade Brasileira de Automática.

Raschka, S. (2024). *Build a Large Language Model (From Scratch)*. Manning Publications Co.

Richards, T. (2023). *Streamlit for Data Science: Create interactive data apps in Python*. Packt Publishing, 2ª edition.

Santos, B. S., Colaço Júnior, M., and Souza, J. G. d. (2018a). An experimental evaluation of the neuromessenger: A collaborative tool to improve the empathy of text interactions. In *2018 IEEE Symposium on Computers and Communications (ISCC)*, pages 00573–00579, Natal. IEEE.

Santos, B. S., Colaço Júnior, M., and Souza, J. G. d. (2018b). A initial experimental evaluation of the neuromessenger: A collaborative tool to improve the empathy of text interactions. In Latifi, S., editor, *Information Technology - New Generations*, pages 411–419, Cham. Springer International Publishing.

Singh, J., Joesph, M. H., and Jabbar, K. B. A. (2019). Rule-based chabot for student enquiries. *Journal of Physics: Conference Series*, 1228(1):012060.

Wohlin, C., Runeson, P., Höst, M., Ohlsson, M. C., Regnell, B., and Wesslén, A. (2024). *Experimentation in software engineering*. Springer Science & Business Media, 2ª edition.