XXXIII Congresso de Iniciação Científica da Unicamp

UM ASSISTENTE CONVERSACIONAL EM AMBIENTE DE APRENDIZAGEM DIGITAL

Palavras-Chave: Retrieval-augmented generation, chatbot, grandes modelos de linguagem

Autores(as):

VICTOR HONÓRIO PEREIRA DA SILVA, IC – UNICAMP

Prof(ª). Dr(ª). JULIO CESAR DOS REIS (orientador), IC – UNICAMP

INTRODUÇÃO

Assistentes conversacionais têm potencial de promover significativas mudanças no meio em que estão inseridos. Dentro do ambiente educacional, por exemplo, podem promover melhorias na aprendizagem dos estudantes e também na manutenção das aulas por parte dos docentes. Todavia, os grandes modelos de linguagem (Large Language Models - LLMs) utilizados atualmente apresentam grandes limitações quando utilizados em contextos específicos de aprendizagem. Nesse sentido, o projeto visa estudar a arquitetura Retrieval-Augmented Generation (RAG) a fim de obter um assistente conversacional que possibilite maior fluidez em um sistema de perguntas e respostas.

Essa proposta de arquitetura possui três componentes principais: um codificador de perguntas, um recuperador de contexto e um gerador de resposta. Conforme esquematizado na Figura 1, o codificador mapeia o texto recebido pela pergunta feita pelo usuário em vetores de embeddings, representando numericamente o valor semântico da informação recebida, dentro de um banco de dados vetorial. O recuperador extrai os candidatos à resposta por meio da busca pelos vetores mais semelhantes à entrada. Por fim, o gerador configura uma resposta dentro dos padrões definidos em linguagem natural para responder a busca codificada.

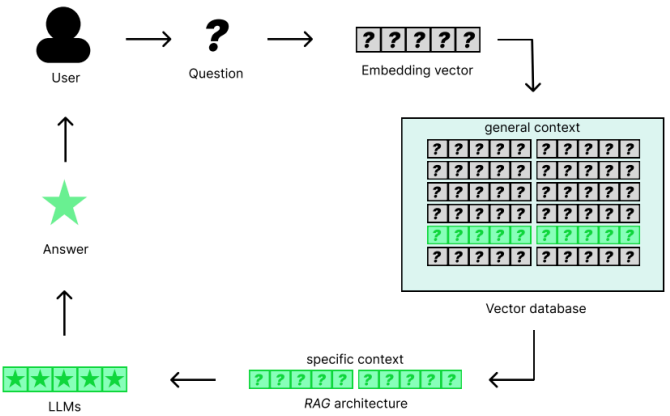

Figura 1: Arquitetura RAG no processo de um assistente conversacional.

METODOLOGIA

Em nossa metodologia, conduzimos primeiramente uma revisão da literatura existente envolvendo a arquitetura RAG e também a utilização de um banco de dados vetorial com embeddings textuais. Foi feita uma análise de tutoriais online e também de artigos para melhor entendimento dos conceitos

envolvendo a pesquisa. Como exemplo, tem-se alguns artigos, o primeiro elucida a importância da arquitetura RAG dentro do escopo de Perguntas e Respostas (Q&A) devido à possibilidade de pesquisa híbrida, tanto acessando a memória interna como buscando externamente o conteúdo da resposta [1], um segundo artigo aborda a utilização da arquitetura RAG para redução de alucinações em sistemas de diálogo [2], o que pode simular uma conversa aluno/professor, corroborando o estudo dessa arquitetura dentro do ambiente educacional e uma terceira fonte importante mostra os desafios encontrados na implementação de um sistema híbrido, haja vista que muitas vezes ocorrem alucinações na resposta gerada pelo chatbot devido a esse conflito que pode ocorrer entre o conhecimento treinado internamente e aquele buscado em fontes externas [3].

Para a escolha do banco de dados, levaram-se em consideração alguns candidatos, dentre os quais o banco Qdrant foi escolhido, devido à sua capacidade de armazenamento na nuvem gratuitamente e também ao fato de ser OpenSource. A partir dos estudos bibliográficos, como próximo passo, foi implementado um primeiro método de assistente conversacional envolvendo a inserção de um arquivo .json no escopo educacional, por meio da criação de uma classe Chatbot e das funções de busca e manutenção da resposta à pergunta feita (query). Inicialmente, era gerado um vetor de *embeddings* para a *query*, passando-se um modelo tokenizador e5-small para obtenção do vetor. Com a *query* vetorizada, era extraído o texto de cada chunk dentro dos resultados e um vetor de *embeddings* para esses textos extraídos era gerado. Tendo o *embedding* da *query* e do *chunk*, era calculada a similaridade entre eles por meio da função *cosine_similarity*, uma função que utilizava o cosseno entre os vetores para verificar o quão alinhados eles estavam, de -1 (opostos) a 1 (perfeitamente alinhados). Com os valores de similaridade dos chunks, era possível ordená-los de acordo com sua proximidade com a *query*, e então retornar uma junção dos chunks mais relevantes dentro do contexto abordado. Por meio desse primeiro estudo, evidenciado na Figura 2 em Resultados, foi possível testar a inserção de dados no banco vetorial e a funcionalidade básica do chatbot, realizando uma busca por uma explicação resumida de uma aula inserida no chatbot.

Dando continuidade aos estudos, optou-se pela criação de pipelines via Google Colab para testar a arquitetura RAG, mantendo-se a utilização do Qdrant como banco de dados vetorial. Primeiramente, foi utilizado o dataset “AGNEWS”, uma base de dados composta por notícias jornalísticas em inglês. Para esta análise, foram selecionados 5 exemplos representativos desse dataset e cada notícia foi adaptada para o formato de pergunta de múltipla escolha, contendo um enunciado, alternativas e um texto de apoio retirado do próprio conteúdo da notícia. Esses dados foram embutidos utilizando o modelo intfloat/multilingual-e5-small e armazenados em uma coleção criada no Qdrant. A arquitetura foi então testada com uma consulta semântica, a partir de perguntas formuladas em inglês e com o objetivo de recuperar o chunk mais relevante entre os vetores armazenados, sendo gerada a partir disso uma tabela com os resultados (Tabela 1).

Um segundo estudo via Colab foi conduzido a partir de um conjunto de dados reais obtidos do “DATASUS”, um dataset referente aos registros de tuberculose extrapulmonar no Brasil entre os anos de 2021 e 2023. Após o tratamento dos dados, foram construídas manualmente 20 perguntas de

múltipla escolha relacionadas ao conteúdo do dataset. Cada pergunta foi associada a um chunk, de forma a alimentar a arquitetura RAG. O objetivo foi verificar a capacidade do componente de recuperação semântica (retriever) em retornar, via busca vetorial, o chunk mais relevante para a pergunta, e então extrair a resposta correta presente no payload associado. O score de similaridade entre o embedding da pergunta e o embedding do chunk recuperado foi analisado, assim como a taxa de acerto Top-1, ou seja, o principal chunk associado a pergunta tema e uma figura foi gerada mostrando os resultados para 10 perguntas dentre essas perguntas criadas (Figura 3).

Por fim, um terceiro estudo foi desenvolvido utilizando algumas transcrições de aulas de uma disciplina do Instituto de Computação da UNICAMP. Para esse estudo, foram processadas cinco aulas no formato .json, contendo textos transcritos das falas do professor. Esses textos foram particionados em *chunks* de até 300 caracteres e armazenados no Qdrant. A partir desse conteúdo, foram elaboradas 10 perguntas relacionadas aos temas discutidos nas aulas, com as suas respectivas respostas esperadas. O sistema foi então testado em dois cenários: um utilizando somente a similaridade do chunk Top-1 recuperado via busca vetorial, e outro combinando os três chunks mais similares (Top-3) como contexto para a geração de uma resposta textual com o modelo *flan-t5-base*. As respostas geradas foram analisadas em termos de similaridade semântica e uma figura foi construída para a análise dos resultados (Figura 4).

RESULTADOS E DISCUSSÃO

A partir da metodologia desenvolvida, foram obtidos alguns resultados evidenciados nas figuras e tabelas abaixo, a detalhar: a Figura 2 mostrando os resultados do primeiro método de implementação; a Tabela 1 indicando os resultados obtidos na análise do primeiro dataset, a Figura 3 detalhando as respostas para o segundo dataset analisado e a Figura 4 mostrando o terceiro dataset.

```
Resposta do chatbot: comentar isso mas mas assim e essa é mais u
ma aula de dúvidas né relacionado ao ao vídeo aula né que foi po
stado mais cedo Uhum mas assim eu por exemplo no meu caso eu não
 consigo assistir esse vídeo antes da aula de noite não sei se s
eria possível postar no final de semana não André a aula não é s
obre que foi a mais cedo não foi a da semana passada ah tá ah be
leza acho que ele caiu ou eu que eu acho que ele travou é é eu t
rouxe dúvida até da aula dois Professor vou deixar anotado pra p
```

Figura 2: primeira resposta obtida ao longo dos estudos

Tabela 1: Resultados obtidos para o primeiro estudo, envolvendo o dataset AGNEWS.

	Pergunta	Resposta	Score
0	Wall St. Bears Claw Back Into the Black (Reuters)	Reuters - Short-sellers, Wall Street's dwindling\\band of ultra-cynics, are seeing green again.	0.78649473

1	Oil and Economy Cloud Stocks' Outlook (Reuters)	Reuters - Soaring crude prices plus worries about the economy and the outlook for earnings are expected to\\hang over the stock market next week during the depth of the summer doldrums.	0.7568784
2	Oil prices soar to all-time record, posing new menace to US economy (AFP)	AFP - Tearaway world oil prices, toppling records and straining wallets, present a new economic menace barely three months before the US presidential elections.	0.7568784

Figura 3: Parte dos resultados obtidos para o segundo estudo, envolvendo o dataset DATASUS.

	Pergunta	Esperado	Retornado	Score	Acerto
0	Qual a região que mais teve casos de tuberculo...	Sul/Sudeste	Sul/Sudeste	0.9495	True
1	Qual agravante teve mais casos de tuberculose ...	HIV	HIV	0.9536	True
2	Qual agravante teve a maior taxa de óbito entr...	HIV	HIV	0.9416	True
3	Qual sexo teve mais casos de tuberculose e qua...	Masculino	Masculino	0.9507	True
4	Qual faixa de idade teve maior número de casos...	26 a 40 anos	26 a 40 anos	0.9574	True
5	Qual faixa de idade teve a maior porcentagem d...	60 anos ou mais	60 anos ou mais	0.9542	True
6	Qual região concentrou o menor número de casos...	Centro-Oeste	Centro-Oeste	0.9583	True
7	Qual comorbidade teve o menor percentual de as...	Doença mental	Doença mental	0.9568	True
8	Em qual faixa etária a tuberculose extrapulmon...	0 a 18 anos	0 a 18 anos	0.9653	True
9	Qual sexo apresentou maior proporção de cura e...	Feminino	Feminino	0.9601	True
10	Qual raça/cor teve a maior taxa de óbito propo...	Parda	Parda	0.9530	True

Figura 4: Parte dos resultados obtidos para o terceiro estudo, envolvendo o dataset MC750.

index	Pergunta	Esperado	Top1 Chunk	Top1 Score
0	O que é Interação Humano-Computador segundo a Aula 01?	Estudo da interação entre pessoas e sistemas computacionais	Humano-Computador Descrever os marcos históricos da área Apresentar perspectivas futuras da área Sistema computacional interativo O que é? Todo sistema computacional é interativo? Exemplos...	0.8924
1	Quais são três impactos negativos da tecnologia mencionados na Aula 01?	Isolamento social, dependência excessiva e exclusão digital	tecnologia, mas vivemos com ela Impactos NEGATIVOS da tecnologia Não apenas usamos tecnologia, mas vivemos com ela Impacto da tecnologia Exemplo: Privacidade Impacto da tecnologia Grandes ...	0.8841
2	O que é um sistema computacional interativo?	Sistema que permite comunicação ativa entre usuário e tecnologia	uso de tecnologia no geral né bom primeiro reflexão que eu trago para vocês e muito relevante é a gente eh refletir e perceber sobre o que significa um sistema computacional interativo né O que é um s...	0.9035
3	Qual exemplo de integração entre vida física e virtual foi citado na Aula 01?	Metaverso	Humano-Computador Descrever os marcos históricos da área Apresentar perspectivas futuras da área Sistema computacional interativo O que é? Todo sistema computacional é interativo? Exemplos...	0.8616
4	O que significa visibilidade no design de interfaces, conforme a Aula 02?	Facilidade de perceber as ações possíveis no sistema	bom primeiro ponto é mantenha elementos da interface sempre visíveis né Quanto mais visíveis é mais provável que os usuários eh vão encontrá-los e saber usá-los Ok por exemplo eh instrumentos no carro...	0.8876
5	Qual a importância do feedback no design de interfaces?	Informar ao usuário sobre o resultado de sua ação	o o curso né e avaliação Então ela é importante para dar feedback paraa equipe de design né se aquelas propostas de design estão indo no caminho certo o que precisa mudar Que tipo de redesign é necess...	0.8849
6	Qual a diferença entre design centrado no usuário e design participativo?	O design participativo envolve o usuário nas decisões de design	decisões de design já no design participativo essa tomada de decisão ela é mais colaborativa né Tom o usuário final Ele também é um elemento que corresponsável na tomada de decisão do Design então ele...	0.9089
7	Segundo a Aula 02, o que é mapeamento?	Relação entre controle e efeito percebido	esse exemplo tá falando do mapeamento mas nesse caso a gente não estaria pensando também num caso de convenção de alguma coisa que é que é convencionada em relação a qual botão acende Qual é pode ter ...	0.8728
8	Na Aula 03, qual conceito refere-se ao grau de tolerância do sistema a erros?	Tolerância a erros	por se nós eh propomos design equivocado de tecnologia e colocarmos isso no mercado né Eh os efeitos só vão ser verificados quando a tecnologia está em operação aí já é tarde demais muito bem bom tudo...	0.8837
9	O que significa consistência no contexto de IHC?	Manutenção de padrões e comportamentos previsíveis	aula A IHC evolui e possui diversos marcos Estabelecida como disciplina em computação Sua evolução desenvolveu um papel chave na adoção e uso de tecnologias mundialmente Referências Baranaus...	0.8693

De acordo com a Figura 2, foram evidenciadas diversas dificuldades no refinamento da resposta tendo em vista a língua portuguesa no tratamento da primeira implementação do Chatbot. Entre os erros, verificaram-se erros de linguagem e truncamentos.

Comparando-se os dois primeiros experimentos realizados via Google Colab — o primeiro com o dataset AGNEWS (Tabela 1) e o segundo com dados do DATASUS (Figura 3) —, é possível observar um aumento considerável no desempenho do sistema em relação ao primeiro Chatbot. No estudo com o dataset de tuberculose, todas as perguntas foram corretamente recuperadas na primeira posição (Top-1), com escores de similaridade acima de 0.93. Esse desempenho indica que o fato das perguntas terem sido formuladas com base nos próprios contextos pode ter facilitado a tarefa do sistema, reduzindo o desafio semântico envolvido. Por outro lado, a linguagem utilizada (português) e a simplicidade das perguntas podem ter contribuído para uma correspondência vetorial mais precisa. Já no estudo com AGNEWS, observou-se uma maior variação nos escores, sugerindo que aspectos como idioma, ruído textual (presença de palavras e caracteres destoantes do restante do contexto salvo no chunk) e o estilo jornalístico podem ter impactado na tarefa de recuperação da informação.

Já no terceiro experimento via Colab (Figura 4), utilizando as transcrições em português, apesar de os escores de similaridade estarem próximos aos obtidos nos estudos anteriores, a geração de respostas apresentou instabilidade, incluindo alucinações e trechos em inglês. Isso sugere que o modelo utilizado para geração (FLAN-T5) tende a responder em sua língua de maior treinamento quando o contexto não está explicitado. Além disso, o formato das instruções no prompt pode ter contribuído para esse comportamento.

CONCLUSÃO

Levando-se em consideração os estudos apresentados até o presente momento, a arquitetura RAG tem demonstrado um avanço no tratamento de sistemas de Perguntas e Respostas, principalmente no tratamento de perguntas diretas, envolvendo alternativas e com o contexto bem definido. No caso de perguntas de construção do conhecimento, perguntas discursivas, o assistente teve variações no rendimento, com o fator da língua podendo ser melhor evidenciado em ocasiões nas quais o assistente ou apresentou vícios de linguagem e truncamentos, dentro de um estudo inicial, ou mesmo optou por realizar a busca e construção da resposta com a sua língua de maior treinamento, evidenciado no estudo final. Todavia, ainda que esses fatores tenham influenciado na construção do assistente, a continuidade dos estudos objetiva reduzir as atuais dificuldades envolvendo a língua portuguesa e contribuir para a construção de um assistente conversacional mais eficaz dentro do ambiente educacional, tanto para questões diretas como para buscas em questões discursivas.

BIBLIOGRAFIA

[1] LEWIS, Patrick, et al. **Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks**. University College London, 2021

[2] SHUSTER, Kurt. **Retrieval Augmentation Reduces Hallucination in Conversation**. Association for Computational Linguistics, 2021

[3] LONGPREE, Shayne. **Entity-Based Knowledge Conflicts in Question Answering**. University of California, Irvine, 2022