



1 INTRODUÇÃO

Este projeto implementa um *chatbot* de perguntas frequentes (FAQ) sobre astronomia observacional, utilizando a técnica de RAG (*Retrieval-Augmented Generation*) sobre um corpus curado de documentos em português.

O sistema recupera trechos relevantes de um conjunto de textos locais e os usa como contexto para um modelo de linguagem gerar respostas fundamentas.

O domínio foi escolhido pois possui perguntas de natureza factual e estável, condição ideal para avaliar um *pipeline* RAG.

Todo o sistema roda localmente e sem custos de API, usando um modelo de código aberto servido pelo Ollama.

2 ARQUITETURA

O sistema se divide em dois fluxos: um de indexação (*offline*, executado uma vez) e um de consulta (*online*, executado a cada pergunta).

2.1 COMPONENTES E TECNOLOGIAS

Camada	Tecnologia	Função
<i>Embeddings</i>	sentence-transformers/all-MiniLM-L6-v2	Converte texto em vetores de 384 dimensões
<i>Vector store</i>	FAISS	Indexa e busca vetores por similaridade
LLM	Mistral-7B-Instruct (quantizado) via Ollama	Gera a resposta a partir do contexto

Orquestração	LangChain (LCEL)	Compõe o <i>pipeline retrieve</i> → <i>generate</i>
Memória	Histórico de mensagens	Permite conversa multi-turno
Interface	CLI em Python e Chainlit	Entrada e Saída de perguntas

3 *CORPUS*

O *corpus* é composto por 15 arquivos de texto em português, cobrindo os principais tópicos de astronomia observacional.

3.1 Estatísticas

Métrica	Valor
Número de documentos	15
Total de palavras	4.633
Total de caracteres	28.359
<i>Chunks</i> gerados (<i>chunk_size</i> =500, <i>overlap</i> =50)	86
Média de caracteres por <i>chunk</i>	329
Dimensão dos vetores	384

3.2 Tópicos

Os tópicos presentes no *corpus* são: Sistema Solar, Lua e fases, constelações, estrelas, galáxias e Via Láctea, buracos negros, telescópios, eclipses, chuvas de meteoros, magnitude, coordenadas celestes, astronomia amadora, observação do céu em São Paulo, história da astronomia e cosmologia.

3.3 Curadoria

Os textos foram redigidos como material educacional conciso, com base em conhecimento consolidado de astronomia, em formato de texto corrido e limpo (sem marcação HTML ou referências), codificados em UTF-8.

4 PIPELINE

4.1 *Chunking*

A divisão em *chunks* usa `RecursiveCharacterTextSplitter`, com `chunk_size=500` caracteres e `chunk_overlap=50`. O *splitter* tenta quebrar primeiro por parágrafos, depois por linhas, frases e palavras, preservando a estrutura semântica. O *corpus* resultou em 86 *chunks*.

4.2 *Embeddings*

Cada *chunk* é convertido em um vetor de 384 dimensões pelo modelo `all-MiniLM-L6-v2`, com normalização dos *embeddings* (torna a similaridade do cosseno equivalente ao produto interno).

4.3 Indexação

Os vetores são indexados no FAISS. Como o *corpus* é pequeno (86 vetores), a busca é exata e ocorre em milissegundos. O índice é persistido em disco e reutilizado nas consultas.

4.4 Recuperação

A cada pergunta, os três *chunks* mais similares são recuperados (`top-k = 3`) por similaridade vetorial.

4.5 Geração

O `Mistral-7B-Instruct`, servido localmente pelo Ollama com temperatura 0,1, recebe os *chunks* recuperados como contexto e gera a resposta. O *prompt* instrui o modelo a usar exclusivamente o contexto fornecido e a recusar perguntas cuja resposta não esteja presente, com exemplos (*few-shot*) do comportamento de recusa.

4.6 Histórico (*bônus*)

Implementou-se um `history-aware retriever`: antes da busca, a pergunta do usuário é reformulada com base no histórico da conversa, de modo que referências como "ele" ou "isso" sejam resolvidas. Por exemplo, após "Me fale sobre Júpiter", a pergunta "Ele tem luas?" é reformulada para "Júpiter tem luas?" antes da busca.

Você: Me fale sobre Júpiter

Bot: Júpiter é o maior planeta do Sistema Solar, com massa maior do que a de todos os outros planetas somados. Ele pertence ao grupo dos planetas gigantes gasosos, compostos principalmente de hidrogênio e hélio. Galileu descobriu as quatro maiores luas de Júpiter no início do século XVII.

Fontes consultadas (3 chunks):

1. 14_historia.txt: No início do século XVII, Galileu Galilei foi um dos primeiros a apontar um telescópio para o céu de...
2. 01_sistema_solar.txt: Júpiter é o maior planeta do Sistema Solar, com massa maior do que a de todos os outros planetas somados...
3. 01_sistema_solar.txt: Os oito planetas, em ordem de distância ao Sol, são: Mercúrio, Vênus, Terra, Marte, Júpiter, Saturno...

Você: Ele é maior que Saturno?

Bot: Sim, Júpiter é maior que Saturno.

Fontes consultadas (3 chunks):

1. 01_sistema_solar.txt: Júpiter é o maior planeta do Sistema Solar, com massa maior do que a de todos os outros planetas somados...
2. 01_sistema_solar.txt: Os oito planetas, em ordem de distância ao Sol, são: Mercúrio, Vênus, Terra, Marte, Júpiter, Saturno...
3. 14_historia.txt: Uma grande mudança ocorreu no século XVI, quando Nicolau Copérnico propôs o modelo heliocêntrico, se...

5 RESULTADO E EVALS

A avaliação usou um conjunto de referência (*golden set*) de 30 perguntas, duas por arquivo do *corpus*, cada uma com resposta-padrão e arquivo-fonte esperado.

5.1 Recuperação: Recall@3

Mede em quantas perguntas o arquivo-fonte esperado apareceu entre os três *chunks* recuperados.

Métrica	Valor
Recall@3	90% (27 de 30)

Análise dos 3 erros de recuperação:

Cada falha ocorreu por um motivo distinto:

Pergunta	Recuperou (incorreto)	Causa provável
"O que são as Três Marias?"	02_lua, 09_chuvas, 01_sistema	A expressão brasileira "Três Marias" não foi bem representada pelo <i>embedder</i> , treinado majoritariamente em inglês
"O que é o Grupo Local?"	09_chuvas, 13_observacao, 12_amadora	Polissemia da palavra "local". A busca casou com arquivos que mencionam "local de observação", em vez do termo técnico "Grupo Local"
"O que Galileu descobriu ao apontar o telescópio?"	07_telescopios, 12_amadora	"Apontar telescópio" casou com o arquivo sobre telescópios, não com o de história.

Os três casos sugerem que *embedder* em inglês é o elo mais frágil do *pipeline* em português, especialmente para expressões idiomáticas, termos técnicos compostos e perguntas cuja intenção difere das palavras exatas usadas na pergunta (*surface words*).

5.2 Geração: Avaliação Subjetiva

Cada resposta gerada foi avaliada manualmente em uma escala de 1 (falha) a 5 (excelente).

Nota	Quantidade	Descrição
5	14	Resposta correta e completa
4	5	Correta, porém resumida
3	4	Parcial, incompleta ou contraditória
2	6	Recusa indevida ou alucinação
1	1	Erro factual
Média		3,83 / 5

6 ANÁLISE DE FALHAS

6.1 Recusa excessiva

Ao reforçar o *prompt* para evitar alucinações, o modelo tornou-se conservador em excesso, recusando perguntas mesmo quando o arquivo correto havia sido recuperado.

Exemplos: "Qual o maior planeta?" e "Quantas constelações?". Em ambos o arquivo-fonte estava no top-3, mas o modelo respondeu: "Não encontrei essa informação."

A causa está na distinção entre granularidades: o Recall@3 mede no nível de arquivo, mas a recuperação acontece no nível de *chunk*. Com 86 *chunks* e $\text{top-k}=3$, o arquivo certo pode aparecer no top-3 sem que o *chunk* específico contendo a sentença-resposta tenha sido recuperado. O modelo recebe contexto do arquivo correto, mas sem a frase exata, e recusa.

Ao aumentar a precisão do *guard-rail* (menos alucinação), aumentou-se a taxa de falsos negativos (mais recusas indevidas).

6.2 Contradição interna

Em alguns casos o modelo recusou e, em seguida respondeu mesmo assim.

Exemplos:

- "Por que Plutão deixou de ser um planeta?": Disse "não encontrei" e depois apresentou uma justificativa incorreta (alegou ser por composição de gelo/poeira. O *corpus* afirma que foi por "não ter limpado a vizinhança da órbita"). Caracteriza vazamento de conhecimento do modelo somado à alucinação.

- "O que são estrelas circumpolares?": "Não encontrei... Porém, sabemos que...", contradiz a própria recusa, embora a resposta final esteja correta.

6.3 Erro Factual

- "Quando ocorre um eclipse solar?": o modelo descreveu o mecanismo do eclipse lunar ("a Lua passa pela sombra da Terra"), trocando os dois fenômenos, embora o *corpus* contenha a definição correta.

- "Tipos de telescópios": listou refrator e refletor, mas afirmou não encontrar o terceiro. O catadióptrico está no *corpus*, mas o *chunk* correspondente não entrou no top-3, resposta incompleta por recuperação parcial.

7 BÔNUS IMPLEMENTADOS

7.1 LangChain (LCEL)

Todo o *pipeline* foi orquestrado com LangChain, usando a composição de *chains* para *retrieve* e *generate*.

7.2 Histórico de conversa

Mantém o histórico e usa um `history-aware retriever` para resolver referências entre turnos.

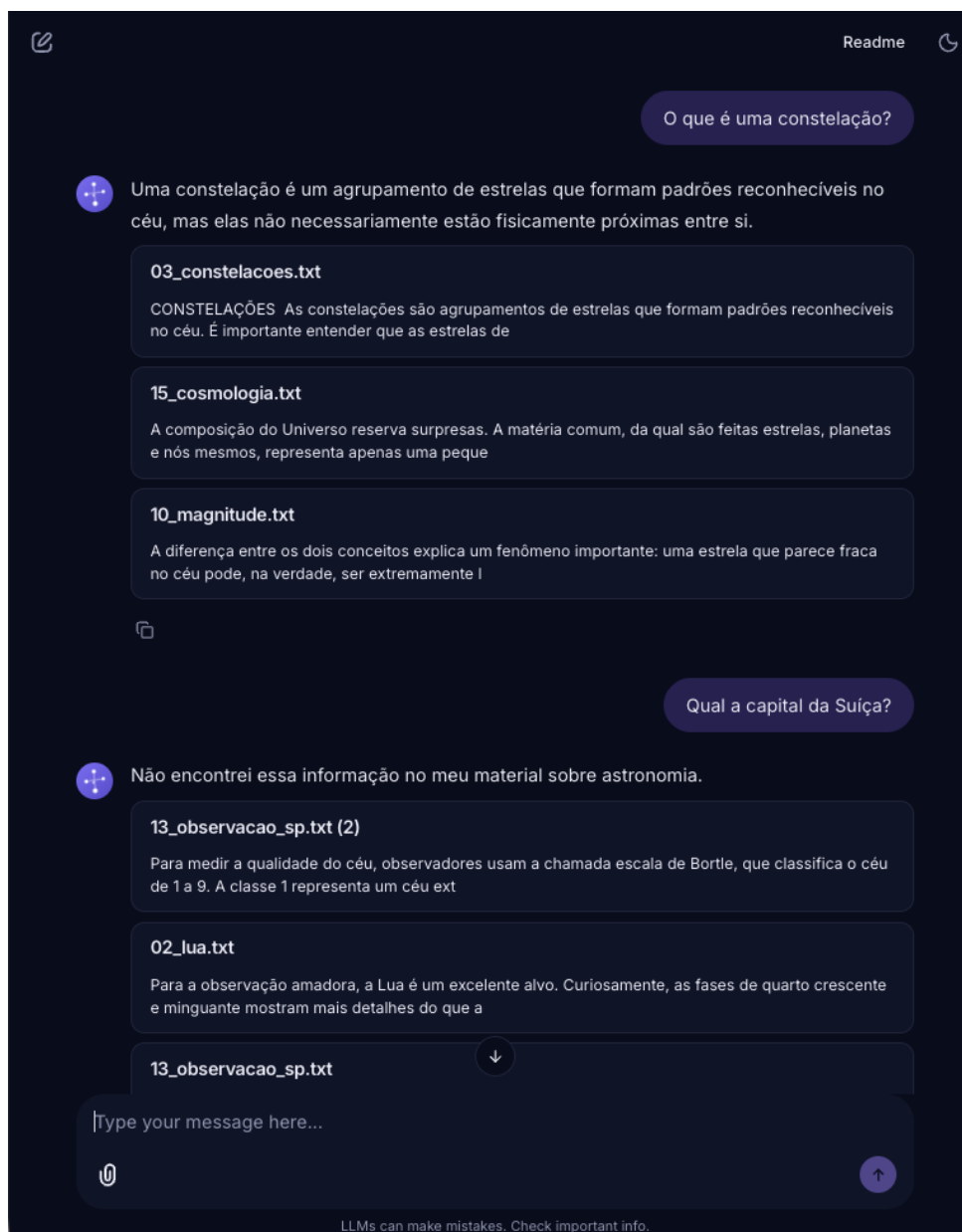
8 IMPLEMENTAÇÃO

O código completo está disponível no GitHub: <https://github.com/cintia-shinoda/astronomia-rag>

Além da interface de linha de comando, o sistema dispõe de uma interface web construída com Chainlit, que oferece uma experiência de chat no navegador, exhibe fontes recuperadas a cada resposta e disponibiliza perguntas-exemplo.

Ambas as interfaces consomem o mesmo pipeline RAG descrito anteriormente.

Captura de tela da interface web:



9 PRÓXIMOS PASSOS

As falhas observadas apontam, todas, para a mesma direção: a evolução natural é de um RAG estático para um *Agentic RAG*, em que o sistema decide ativamente como recuperar e quando responder.

Melhorias:

1. ***Grading* de relevância:** avalia se os *chunks* recuperados realmente respondem à pergunta antes de gerar. Atacaria o *over-refusal* e o uso de contexto parcial.
2. ***Query rewriting*/expansão:** reformular ou expandir a pergunta antes da busca, mitigando falhas como "Três Marias" e "Grupo Local".
3. ***Embedder* multilíngue:** substituir o `all-MiniLM-L6-v2` por um modelo treinado também em português (por exemplo da família *paraphrase-multilingual*).
4. **Aumento de *top-k* ou *re-ranking*:** Recuperar mais *chunks* e reordená-los, reduzindo o problema de granularidade arquivo vs. *chunk*.

Essas extensões transformam o pipeline reativo atual em um agente que planeja a recuperação.

10 REFERÊNCIAS

Bibliotecas e ferramentas:

- LangChain: python.langchain.com
- FAISS (*Facebook AI Similarity Search*): github.com/facebookresearch/faiss
- Sentence-Transformers: [sbert.net](https://www.sbert.net)
- Ollama: ollama.com
- Mistral-7B-Instruct: Mistral AI

Corpus:

Os 15 documentos do corpus foram redigidos com auxílio de assistente de IA, a partir de conhecimento consolidado de astronomia. As fontes abaixo embasam e permitem a verificação do conteúdo:

Oliveira Filho, K. S.; Saraiva, M. F. O. Astronomia e Astrofísica. Universidade Federal do Rio Grande do Sul (UFRGS). Disponível em:

<http://astro.if.ufrgs.br/livro.pdf>

Wikipédia em português (CC BY-SA 4.0): artigos correspondentes a cada tópico (Sistema Solar, Lua, Constelação, Estrela, Via Láctea, Buraco negro, Telescópio, Eclipse, Chuva de meteoros, Magnitude aparente, Coordenadas celestes, Astronomia amadora, Poluição luminosa, História da astronomia, Cosmologia).

Observatório Nacional (ON/MCTI): <https://www.gov.br/observatorio>

União Astronômica Internacional (IAU): <https://www.iau.org>