

LangChain: Criando chatbots inteligentes com RAG

Avaliação com LangSmith & RAGAS

Instrutor(a): Leonardo Pena

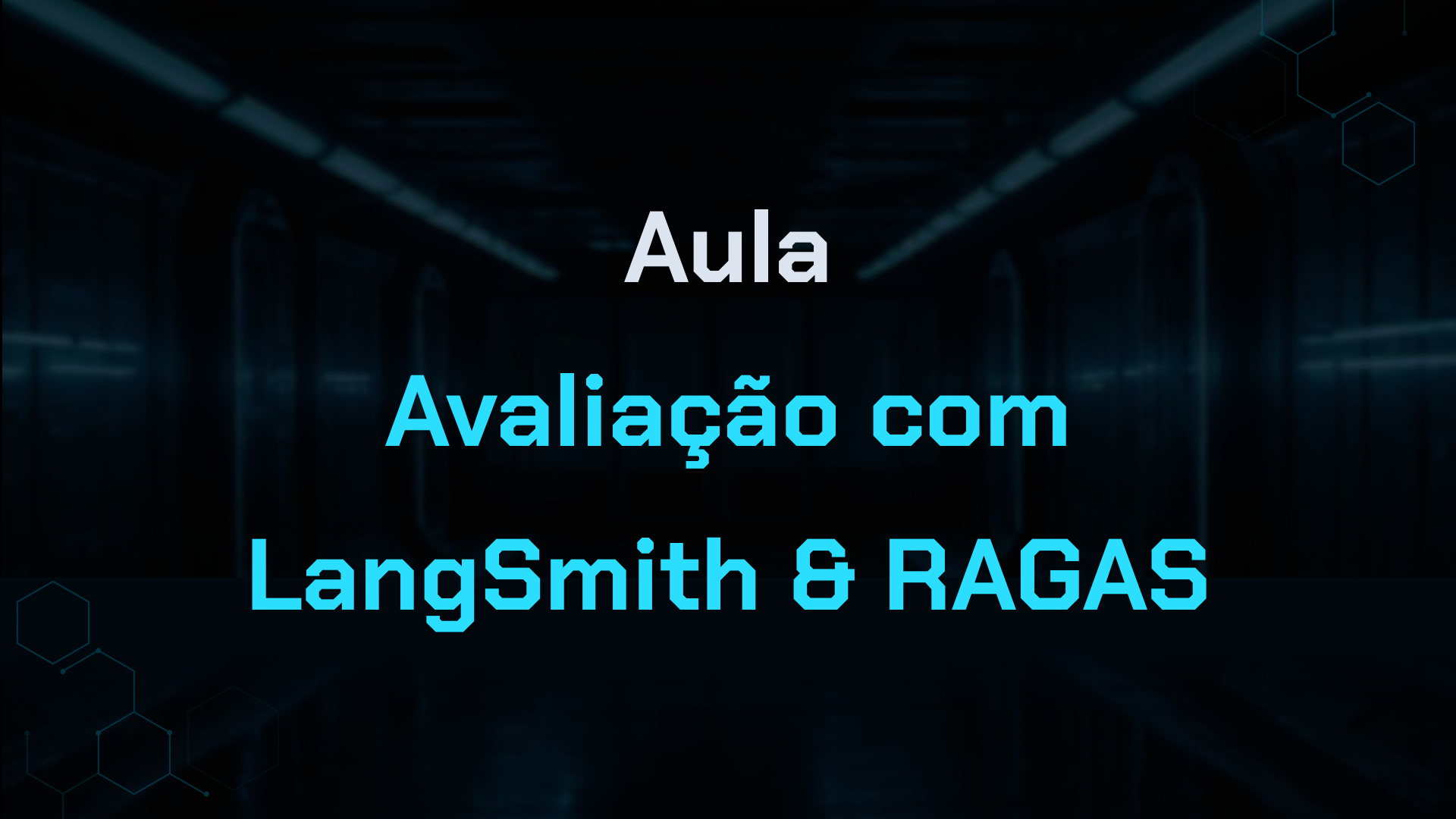


MERGULHE EM TECNOLOGIA_



Seja bem vindo (a)!





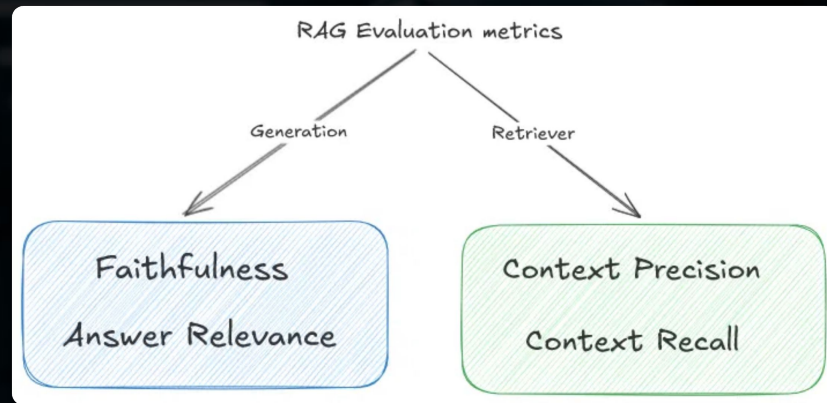
Aula

Avaliação com

LangSmith & RAGAS

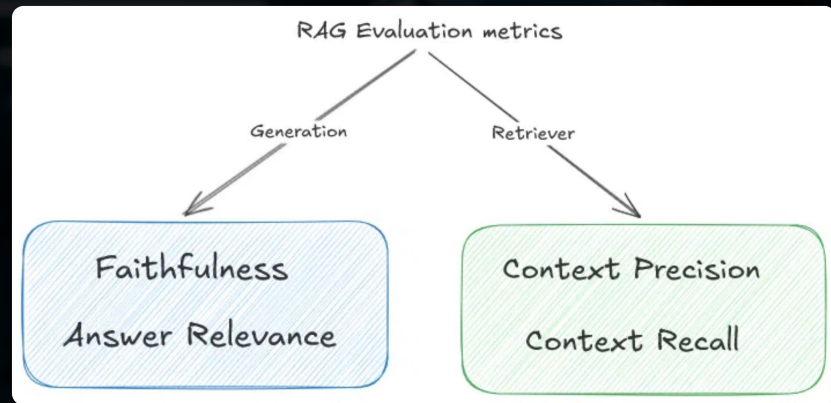
Desafios na Avaliação de RAG

Avaliar sistemas RAG é **significativamente mais complexo** do que avaliar LLMs tradicionais, devido à natureza multifacetada desses sistemas.



Desafios na Avaliação de RAG

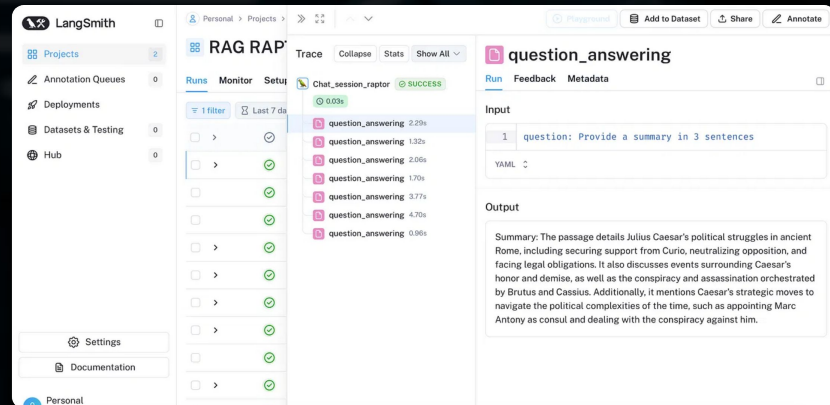
- ≡ Múltiplas camadas para avaliar: recuperação de documentos e geração de respostas
- ≡ Dificuldade em automatizar a verificação da factualidade das respostas
- ⚖ Subjetividade na avaliação da relevância das respostas e do contexto
- Interdependência entre componentes:
 - 🧩 falhas em uma etapa afetam as subsequentes



Esses desafios exigem abordagens especializadas e ferramentas dedicadas para avaliação eficaz.

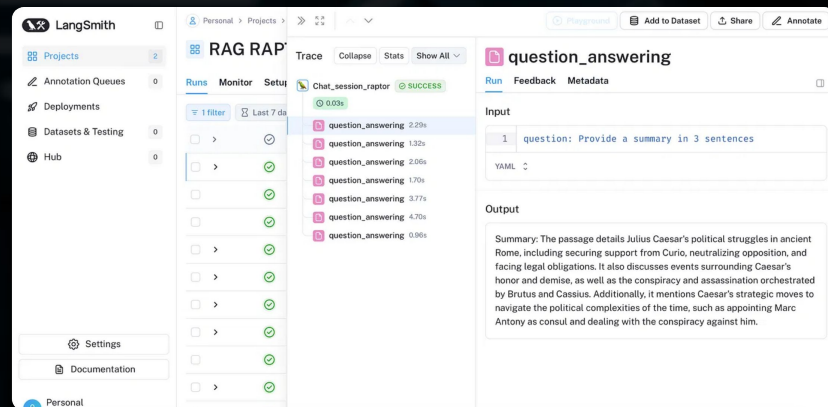
LangSmith: Plataforma de Observabilidade

LangSmith é uma plataforma de observabilidade desenvolvida pela LangChain, projetada para depurar, testar, avaliar e monitorar aplicações de LLMs.



LangSmith: Plataforma de Observabilidade

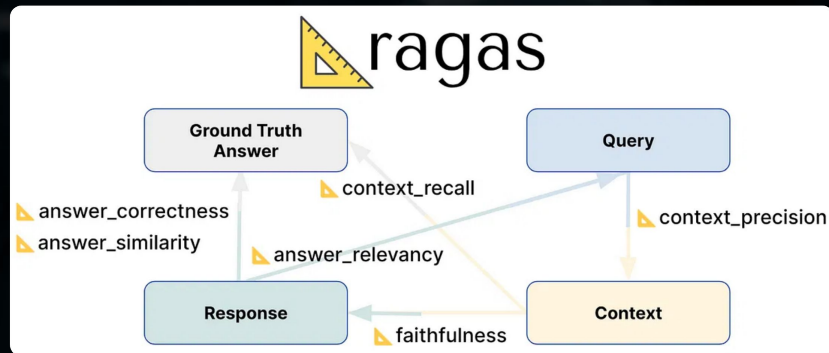
- 🔍 Rastreamento (tracing) de cada passo da execução de uma cadeia
- 🛑 Identificação de gargalos e erros em tempo real
- 📈 Monitoramento de métricas de desempenho e custos
- 🔧 Ambiente para testes e experimentação com diferentes configurações



A plataforma permite visualizar todo o fluxo de execução, facilitando a compreensão de como cada componente contribui para o resultado final.

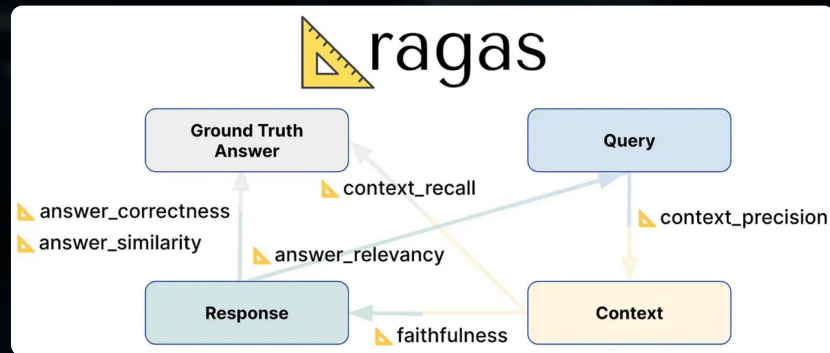
RAGAS: Biblioteca de Avaliação

RAGAS (Retrieval Augmented Generation Assessment) é uma biblioteca de código aberto que fornece métricas específicas para avaliar a performance de sistemas RAG.



RAGAS: Biblioteca de Avaliação

- 🎯 Foco na qualidade da interação entre recuperação e geração
- 📋 Métricas específicas para sistemas RAG, além das métricas tradicionais
- 🔗 Fácil integração com pipelines existentes e frameworks como LangChain
- 🤖 Avaliação automatizada usando LLMs para reduzir a necessidade de avaliação humana



O RAGAS permite uma **avaliação holística** dos sistemas RAG, considerando tanto a qualidade da recuperação quanto a precisão da geração.

Métricas de Geração

O RAGAS fornece métricas específicas para avaliar a **qualidade da geração** de respostas em sistemas RAG, focando em dois aspectos principais:

ragas score

generation

faithfulness

how factually accurate is
the generated answer

answer relevancy

how relevant is the generated
answer to the question

retrieval

context precision

the signal to noise ratio of retrieved
context

context recall

can it retrieve all the relevant information
required to answer the question

Métricas de Geração



Factualidade (Faithfulness)

Mede se a resposta gerada é suportada pelos documentos recuperados, evitando alucinações.
Exemplo: Se a resposta contém informações que não estão presentes nos documentos recuperados, a pontuação de factualidade será baixa.



Relevância da Resposta (Answer Relevance)

Avalia se a resposta gerada é relevante para a pergunta feita pelo usuário.
Exemplo: Uma resposta pode ser factualmente correta, mas não responder diretamente à pergunta do usuário, resultando em baixa relevância.

ragas score

generation

faithfulness

how factually accurate is
the generated answer

answer relevancy

how relevant is the generated
answer to the question

retrieval

context precision

the signal to noise ratio of retrieved
context

context recall

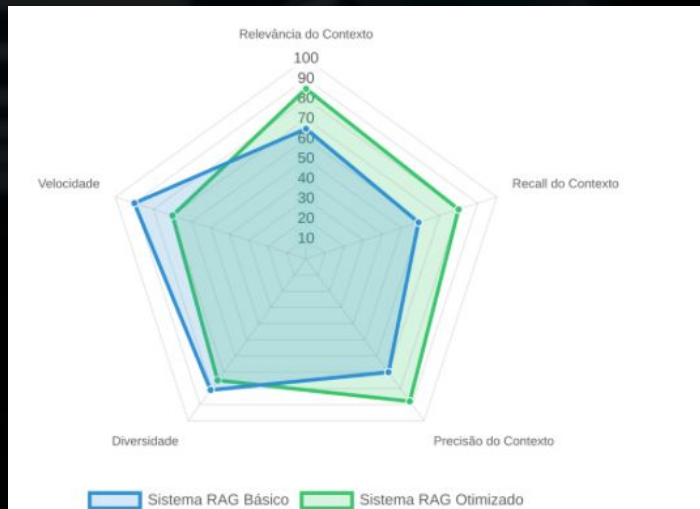
can it retrieve all the relevant information
required to answer the question

Estas métricas são complementares e devem ser analisadas em conjunto para uma avaliação completa da qualidade da geração.

Métricas de Recuperação

As métricas de recuperação do RAGAS avaliam a qualidade dos documentos recuperados pelo sistema RAG, garantindo que o contexto fornecido ao LLM seja relevante e completo.

Comparação das Métricas de Recuperação



Métricas de Recuperação



Relevância do Contexto (Context Relevance)

Avalia se os documentos recuperados são pertinentes para a consulta do usuário, medindo a similaridade semântica entre a consulta e cada documento.



Recall do Contexto (Context Recall)

Mede se todos os documentos necessários para responder à consulta foram recuperados, avaliando a cobertura das informações relevantes.



Precisão do Contexto (Context Precision)

Avalia a proporção de documentos recuperados que são realmente relevantes, penalizando a inclusão de informações irrelevantes.

Estas métricas são complementares e devem ser analisadas em conjunto para uma avaliação completa da etapa de recuperação.

Comparação das Métricas de Recuperação

