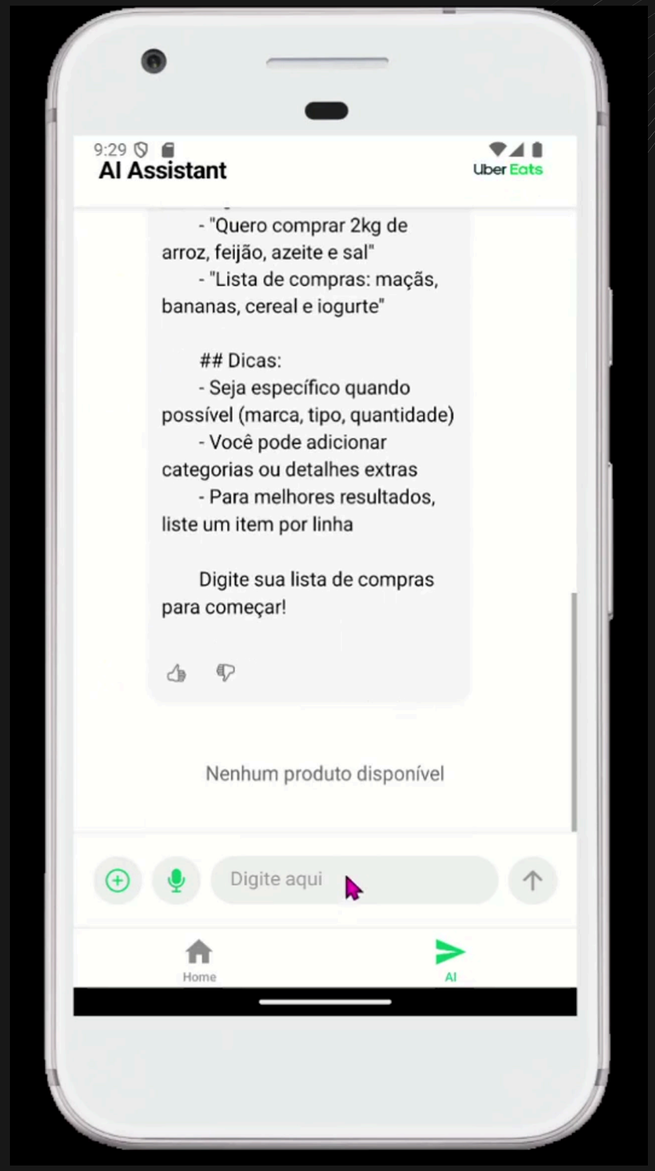
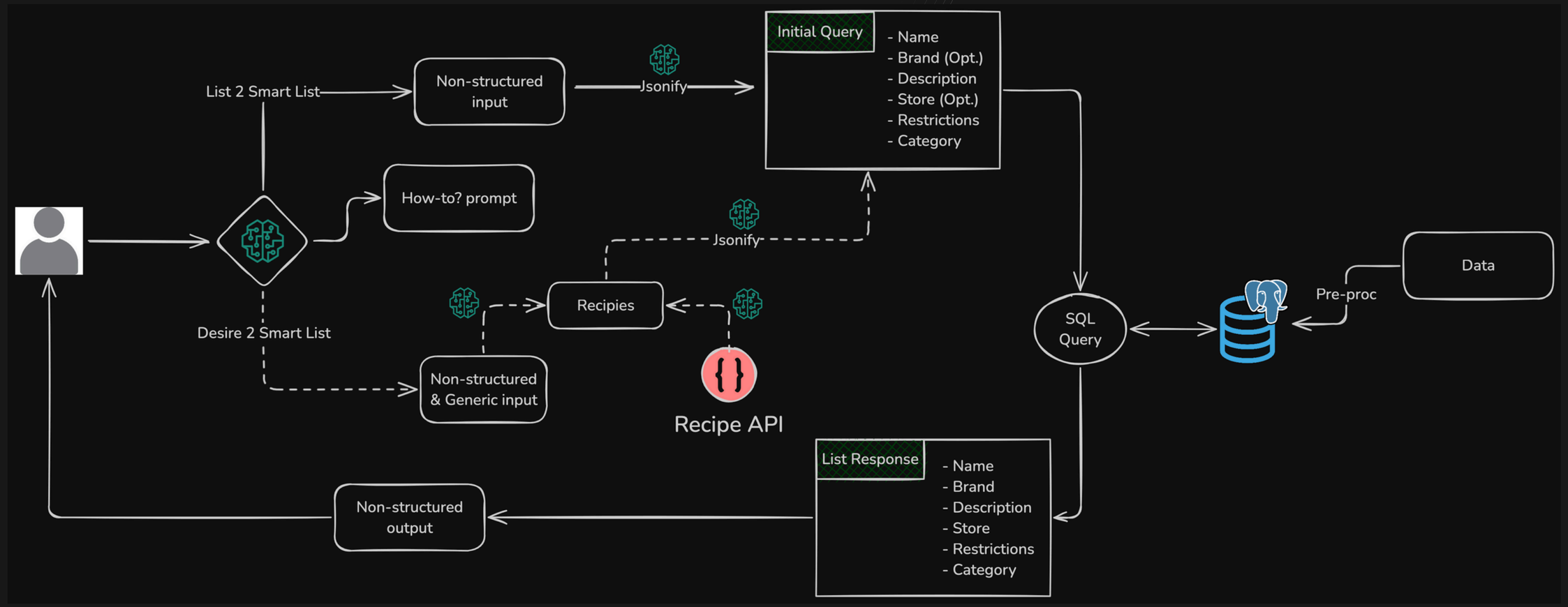
UberChat

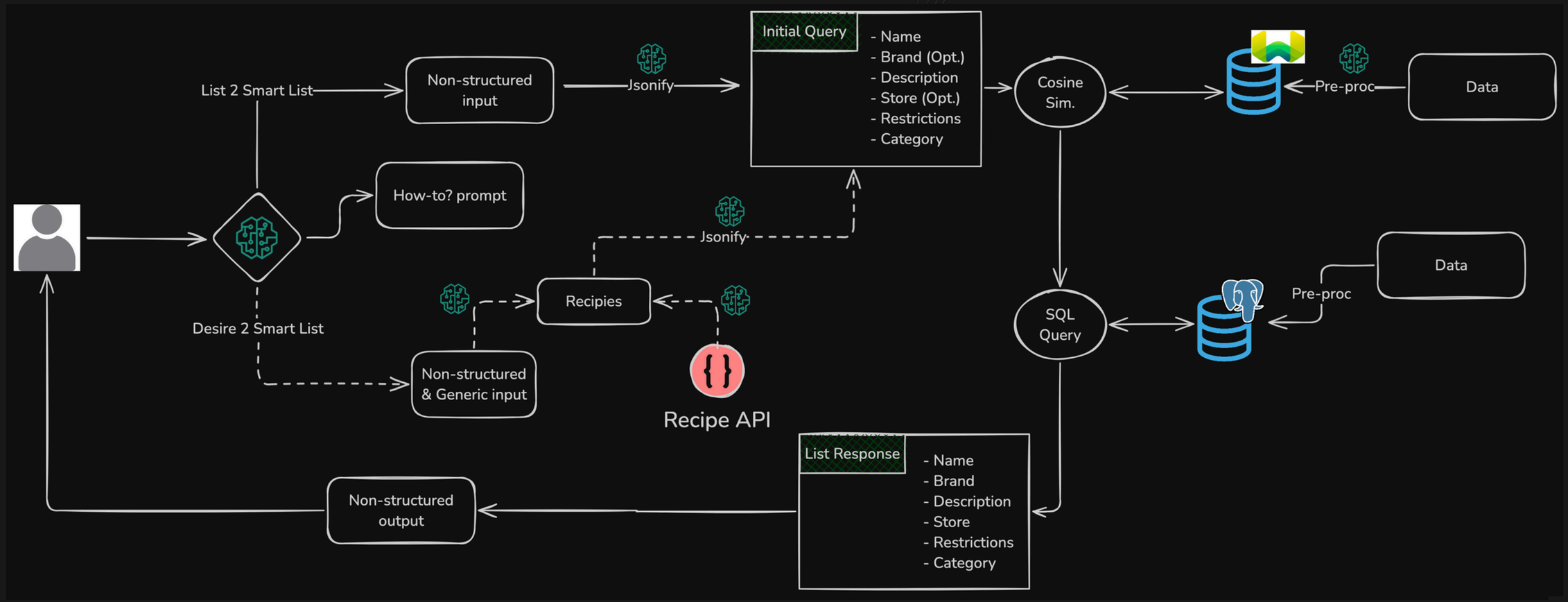
Sprint 03


9:29
AI Assistant
Uber Eats
- "Quero comprar 2kg de arroz, feijão, azeite e sal"
- "Lista de compras: maçãs, bananas, cereal e iogurte"
Dicas:
- Seja específico quando possível (marca, tipo, quantidade)
- Você pode adicionar categorias ou detalhes extras
- Para melhores resultados, liste um item por linha
Digite sua lista de compras para começar!
Nenhum produto disponível
Digite aqui
Home
AI

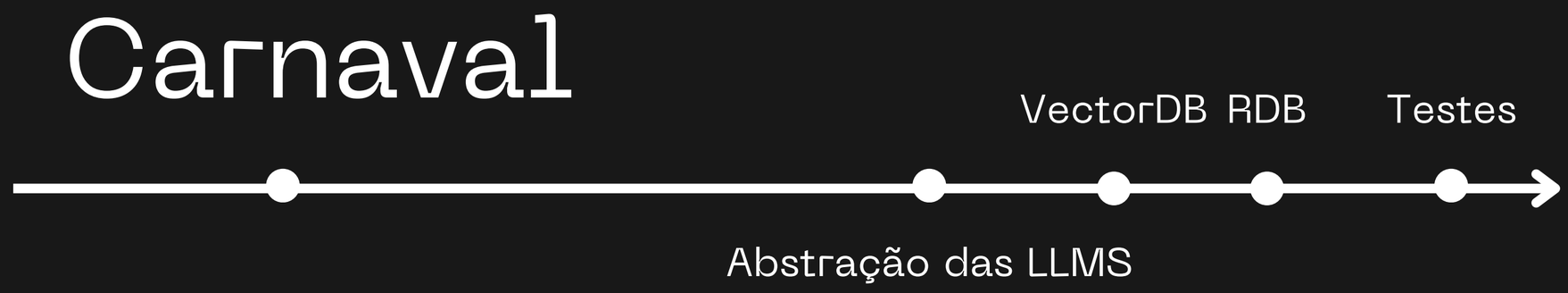
Sprint 2


List 2 Smart List
Non-structured input
Jsonify
Initial Query
- Name
- Brand (Opt.)
- Description
- Store (Opt.)
- Restrictions
- Category
How-to? prompt
Jsonify
Data
Pre-proc
Recipies
SQL Query
Desire 2 Smart List
Non-structured & Generic input
Recipe API
List Response
- Name
- Brand
- Description
- Store
- Restrictions
- Category
Non-structured output

Sprint 2


List 2 Smart List
Non-structured input
Jsonify
Initial Query
- Name
- Brand (Opt.)
- Description
- Store (Opt.)
- Restrictions
- Category
Cosine Sim.
Pre-proc
Data
How-to? prompt
Jsonify
Recipies
Desire 2 Smart List
Non-structured & Generic input
Recipe API
SQL Query
Pre-proc
Data
List Response
- Name
- Brand
- Description
- Store
- Restrictions
- Category
Non-structured output

Nossa sprint



Preprocessamento com LLM

Extração de Informações Relevante

Normalização de Dados

Geração de Resumos ou Metadados

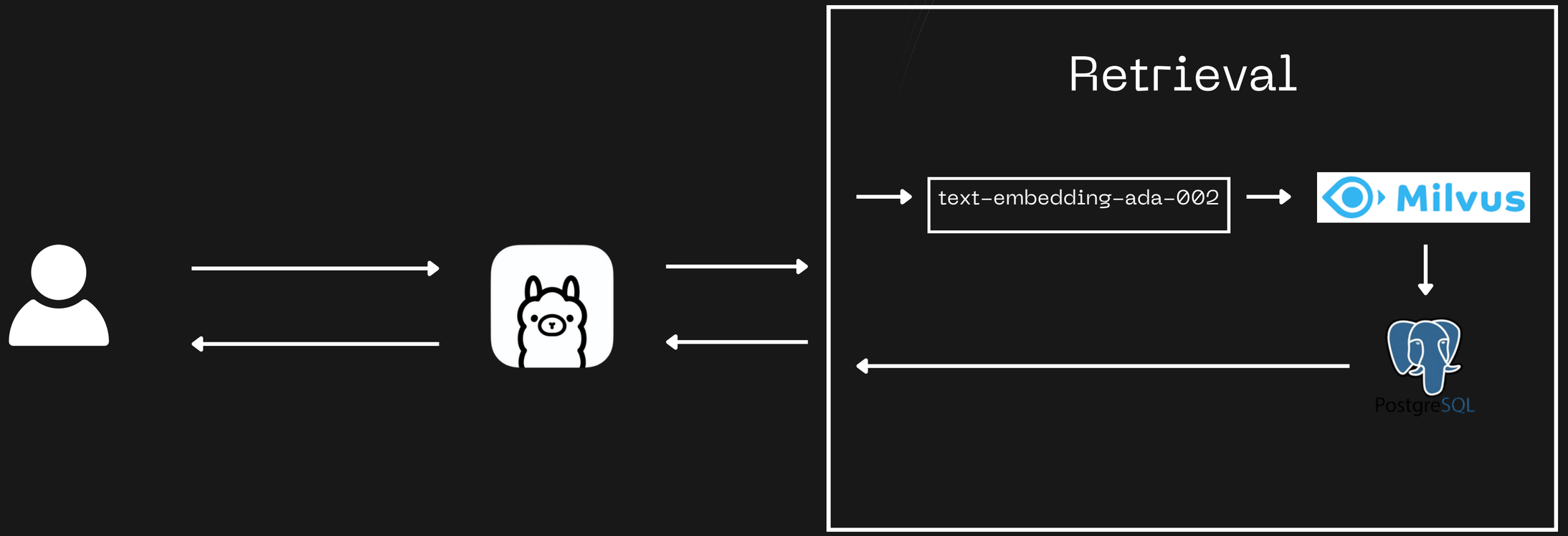
RAG

RAG

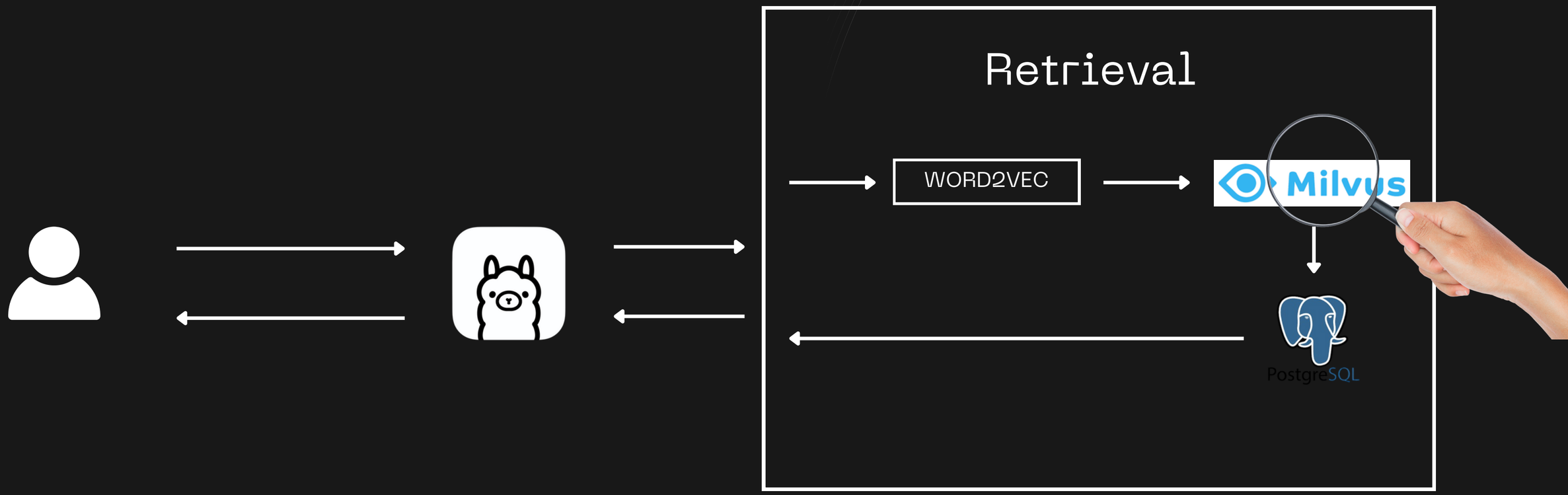
- Geração de resposta adaptada para o problema
- Reduz o viés e as "alucinações" dos modelos geradores.
- Melhora a precisão e a qualidade das respostas
- Redução de custos computacionais

IMPLEMENTAÇÃO

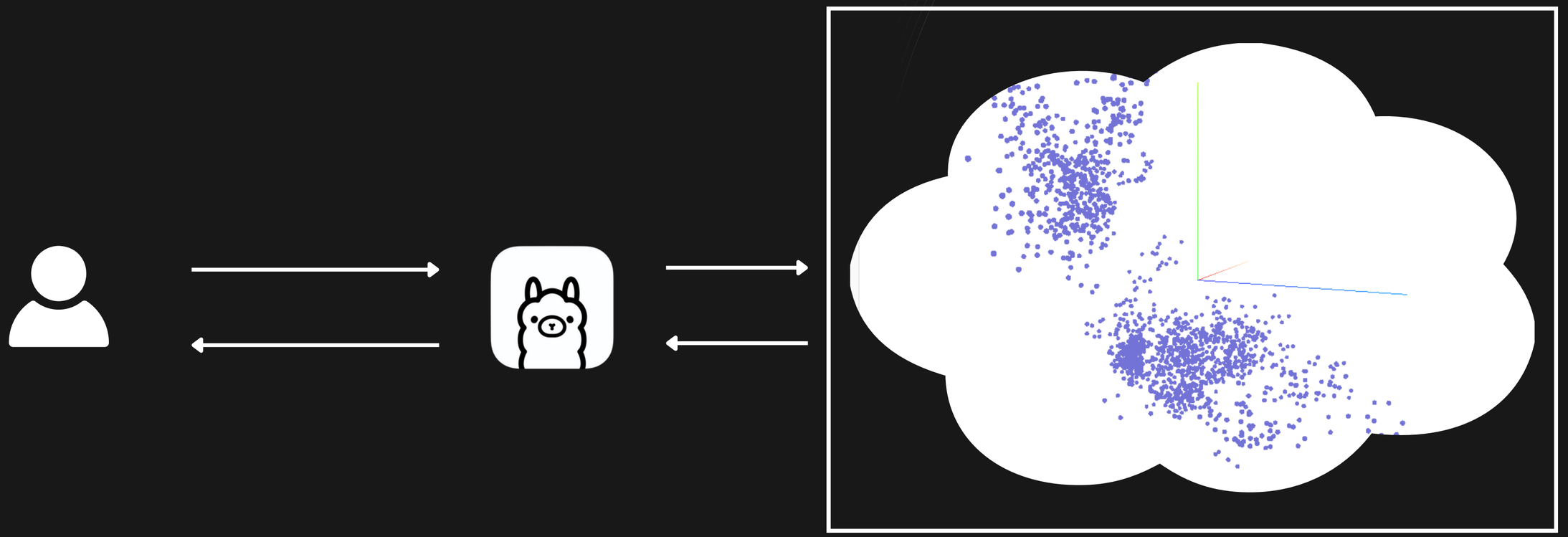
RAG



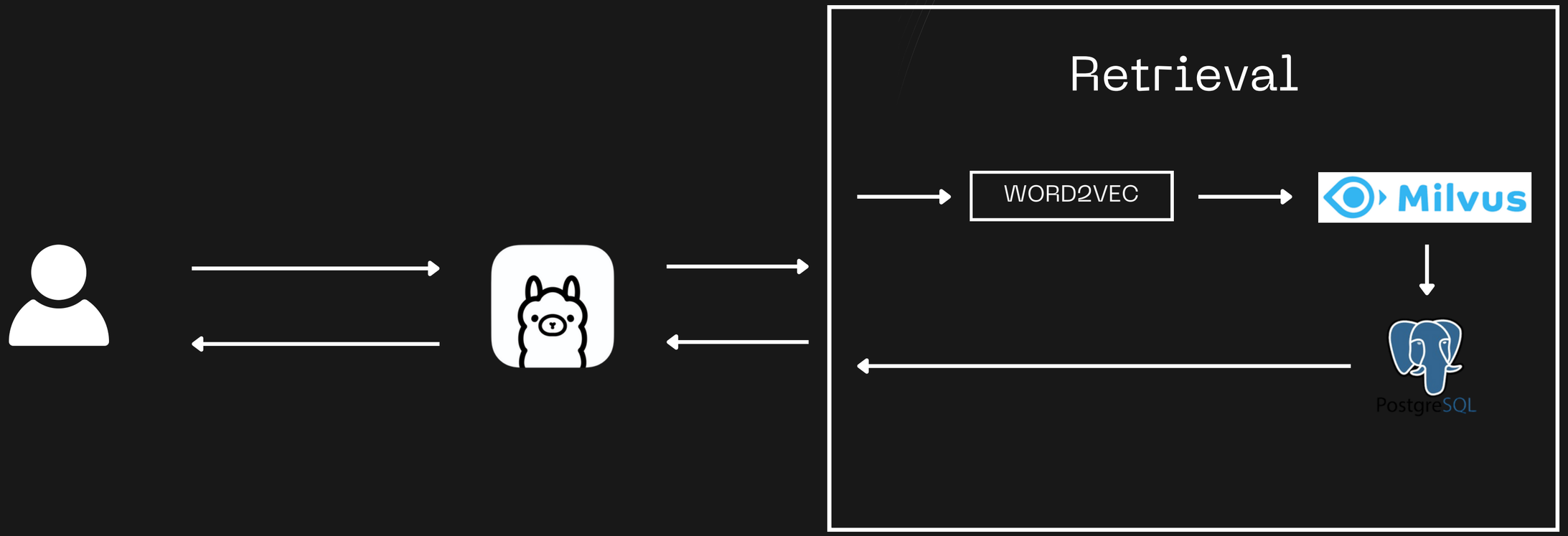
RAG



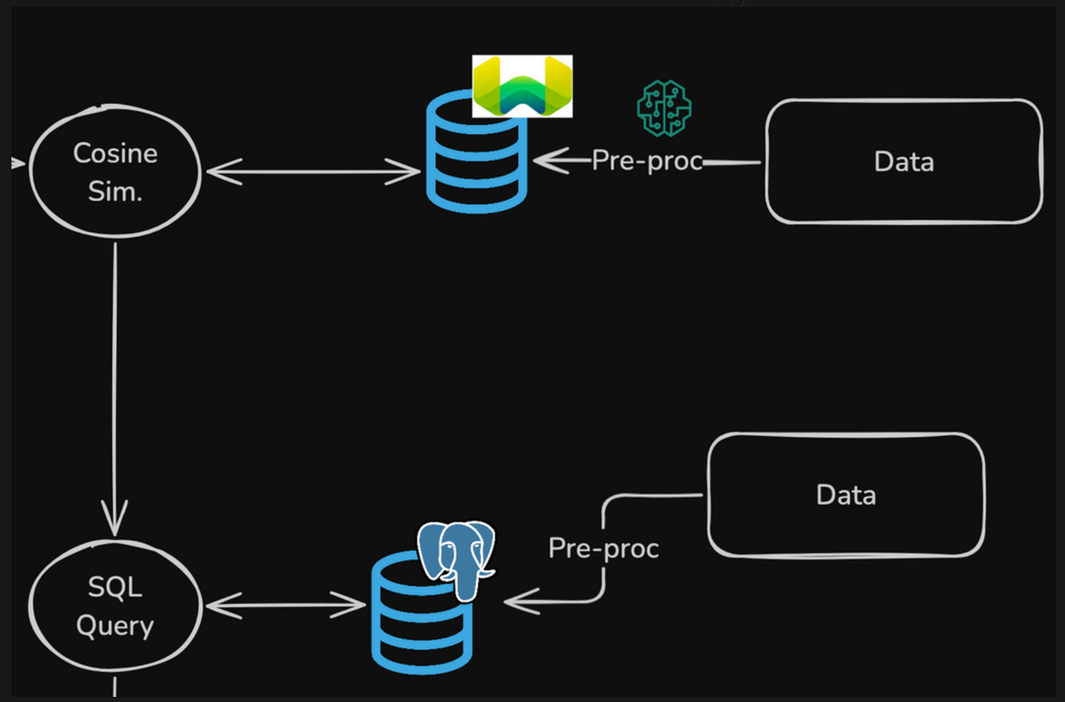
RAG



RAG



RAG no fluxo da aplicação



Testes

Como validamos nosso sistema?

Testes de Recuperação (test_retrieval.py)

. Está retornando os documentos corretos?

. Simulando buscas reais

Testes de Geração (test_generation.py)

. Consegue produzir respostas coerentes?

. Comparando com respostas esperadas

Testes de Integração (test_integration.py)

. O fluxo funciona?.

. Simulando um fluxo completo: buscar + responder

RAGAS

Testes

RAGAS

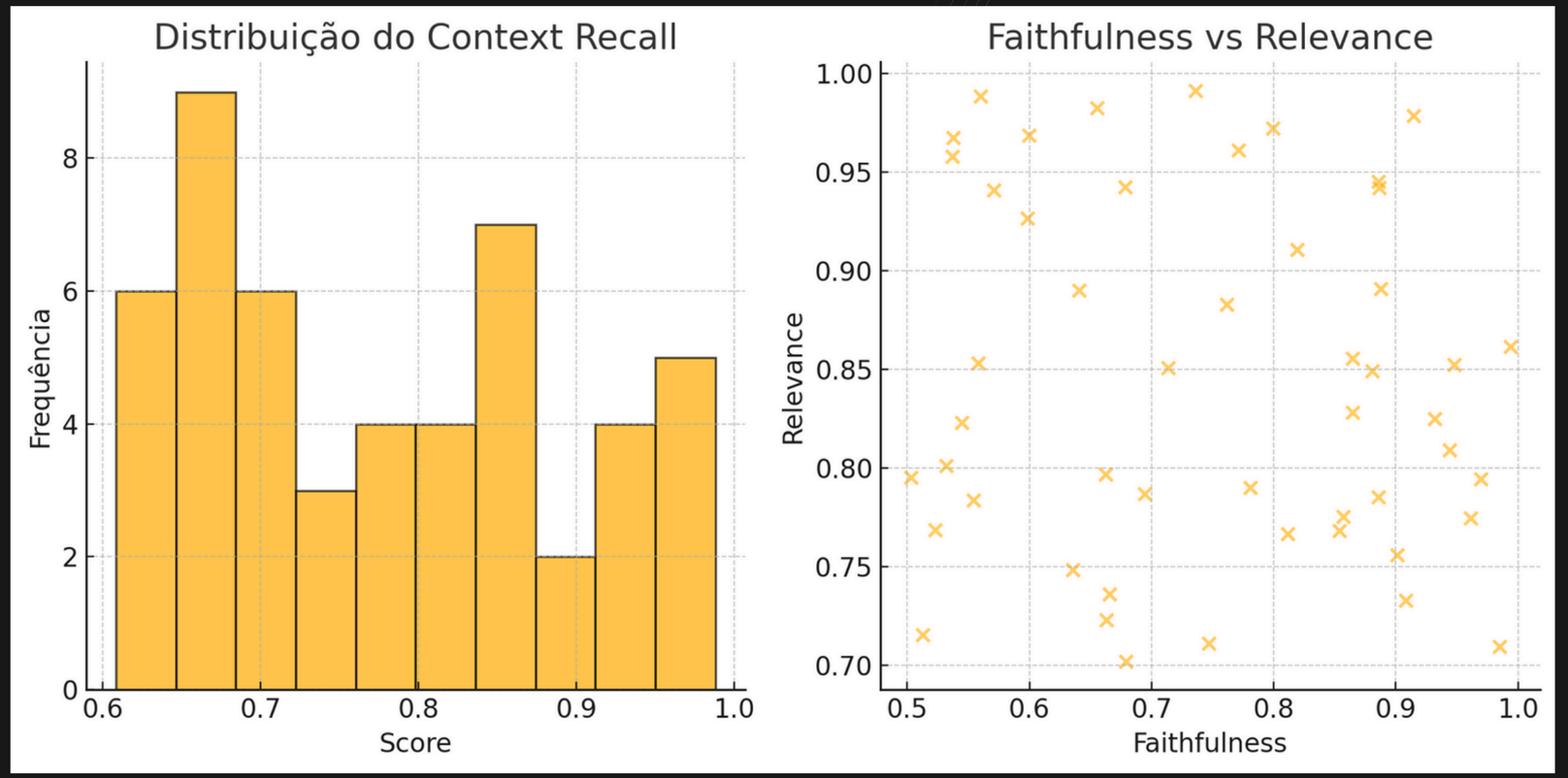
Relevância	Precisão	Coerência
Fidelidade da geração	Precisão da recuperação	Relevância do contexto

Resultados

#	context_recall	faithfulness	relevance
	749816047538945	9.847.923.138.822.790	7.094.287.557.060.200
	9,80286E+15	8.875.664.116.805.570	8.909.231.233.791.340
	892797576724562	9.697.494.707.820.940	794.306.794.322.898
	8,39463E+15	9.474.136.752.138.240	8.525.712.073.494.100
	6,62407E+15	7.989.499.894.055.420	9.722.699.421.778.270
	662397808134481	9.609.371.175.115.580	7.747.876.687.446.620
	6,23233E+15	5.442.462.510.259.590	8.231.148.769.106.880
	9,4647E+15	5.979.914.312.095.720	9.266.653.415.629.140
	8,40446E+15	522.613.644.455.269	7.686.394.496.474.860
	8,83229E+15	6.626.651.653.816.320	7.230.939.729.486.370
	608233797718321	6.943.386.448.447.410	7.869.254.358.741.300
	9,87964E+15	6.356.745.158.869.470	7.483.663.861.762.010
	9,32977E+15	9.143.687.545.759.640	9.789.092.957.027.710
	6,84936E+15	6.783.766.633.467.940	9.424.361.138.693.250
	6,7273E+15	6.404.672.548.436.900	8.900.211.269.531.270
	6,73362E+15	7.713.480.415.791.240	9.614.381.770.563.150
	7,21697E+15	5.704.621.124.873.810	9.411.016.230.697.340
	8,09903E+15	9.010.984.903.770.190	7.559.710.176.658.100
	7,72778E+15	5.372.753.218.398.850	9.677.676.995.469.930
	7,16492E+15	9.934.434.683.002.580	8.618.026.725.746.950

#	context_recall	faithfulness	relevance
	8,44741E+15	8.861.223.846.483.280	9.422.320.465.492.180
	6,55798E+15	5.993.578.407.670.860	968.827.389.977.048
	7,16858E+15	5.027.610.585.618.010	7.954.010.424.915.590
	7,46545E+15	9.077.307.142.274.170	733.015.577.358.303
	7,82428E+15	8.534.286.719.238.080	7.683.805.487.625.820
	9,1407E+15	8.645.035.840.204.930	8.281.323.365.878.760
	6,7987E+15	8.856.351.733.429.720	9.454.044.297.767.470
	8,05694E+15	5.370.223.258.670.450	958.219.174.976.903
	836965827544817	6.792.328.642.721.360	7.020.856.391.593.570
	6,1858E+15	5.579.345.297.625.640	8.532.241.907.732.690
	8,43018E+15	9.315.517.129.377.960	8.252.233.009.446.330
	6,6821E+15	811.649.063.413.779	7.666.323.431.412.190
	6,26021E+15	6.654.490.124.263.240	7.359.596.102.001.040
	9,79554E+15	5.317.791.751.430.110	8.012.845.514.210.880
	9,86253E+15	6.554.911.608.578.310	9.828.729.111.737.550
	9,23359E+15	6.625.916.610.133.730	7.969.608.796.062.260
	7,21846E+15	864.803.089.169.032	8.556.371.865.230.090
	6,39069E+15	8.187.787.356.776.060	9.109.056.876.685.530
	8,73693E+15	9.436.063.712.881.630	8.090.888.807.137.880
	7,76061E+15	7.361.074.625.809.740	9.915.346.248.162.880

Resultados

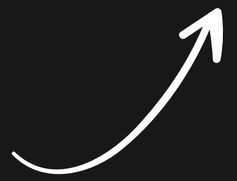

Distribuição do Context Recall
Frequência
Score
0
2
4
6
8
0.6
0.7
0.8
0.9
1.0
Faithfulness vs Relevance
Relevance
Faithfulness
0.70
0.75
0.80
0.85
0.90
0.95
1.00
0.5
0.6
0.7
0.8
0.9
1.0

Testes

Por que esses resultados são importantes?

Implementação de métricas de avaliação contínua.

Refinamento do modelo e da base de dados.



Respostas coerentes e úteis para os usuários.

RERANKER

Mais Precisão: Ajusta a ordem das respostas para resultados mais exatos.

Aplicações: Perfeito para busca ou recomendações onde relevância importa.

Flexibilidade: Personaliza critérios como contexto ou confiabilidade.

DEMO

Observações

Contexto da Lista de Compras

- Lista de compras – genérico
- Problema para similaridade

Processo com LLM

- Resumir X Gerar
- Busca por similaridade

Observações — Problemas Identificados

Geração de frases aleatórias

Problemas na busca

O caso da pasta de dente

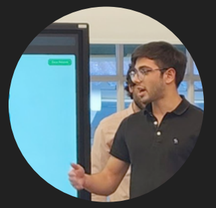
Próximos Passos

Desenvolver um **teste End-to-End** que permita avaliar como uma **alteração** em nossa abordagem afeta o **resultado final**, proporcionando uma compreensão clara do impacto das mudanças realizadas.

Meet our team

Marcos Teixeira

Fabio Piemonte



Guilherme Lima

Beny Frid

Enya Oliveira

Perguntas ?