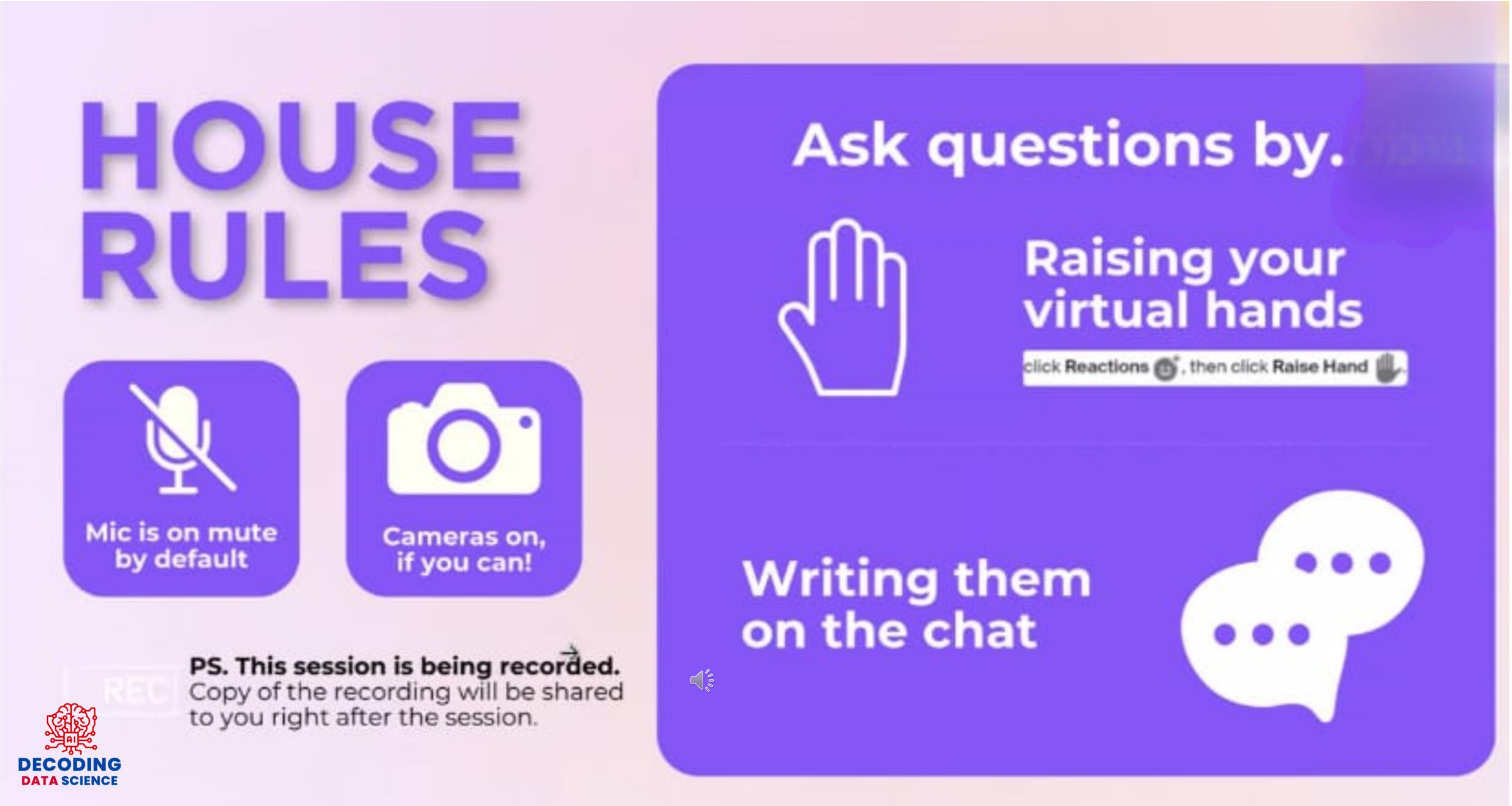


HOUSE RULES

Mic is on mute by default

Cameras on, if you can!

PS. This session is being recorded.
Copy of the recording will be shared to you right after the session.

REC

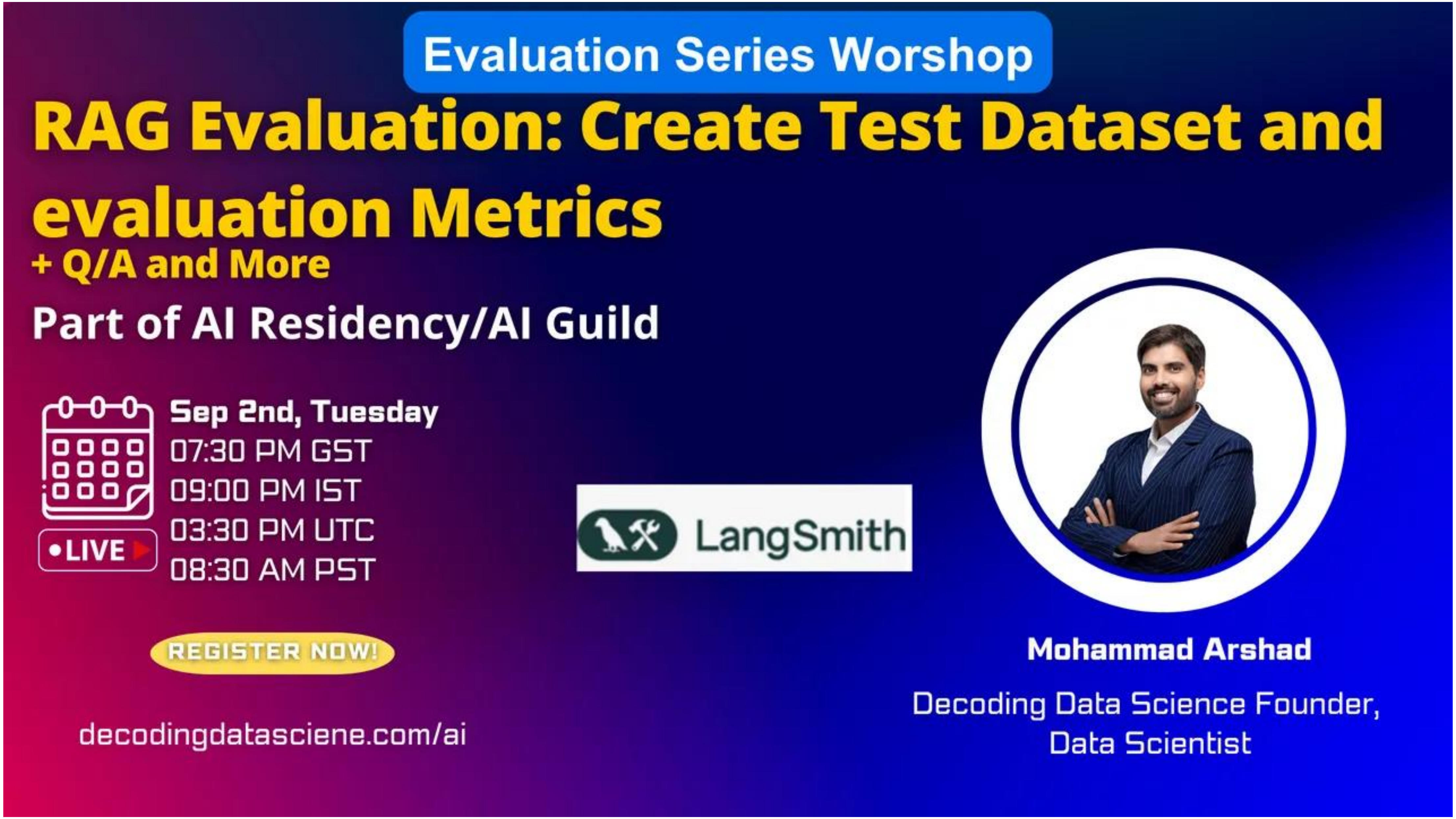
Ask questions by.

Raising your virtual hands

click **Reactions**, then click **Raise Hand**

Writing them on the chat

DECODING
DATA SCIENCE



Evaluation Series Worshop

RAG Evaluation: Create Test Dataset and evaluation Metrics

+ Q/A and More

Part of AI Residency/AI Guild

Sep 2nd, Tuesday

07:30 PM GST
09:00 PM IST
03:30 PM UTC
08:30 AM PST

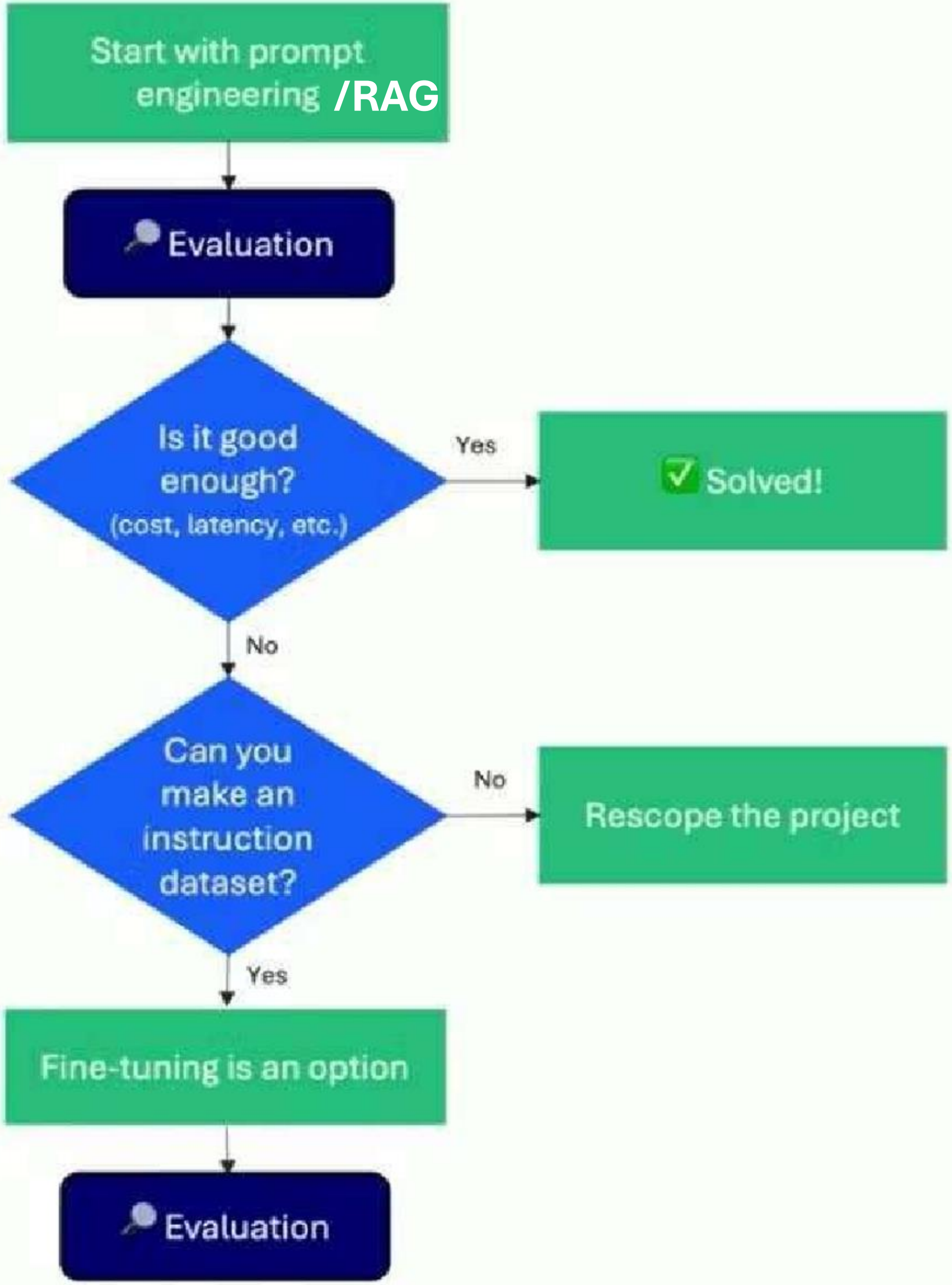
•LIVE

REGISTER NOW!

decodingdatasciene.com/ai

Mohammad Arshad

Decoding Data Science Founder,
Data Scientist



Start with prompt engineering /RAG

🔎 Evaluation

Is it good enough? (cost, latency, etc.)

Yes

✅ Solved!

No

Can you make an instruction dataset?

No

Rescope the project

Yes

Fine-tuning is an option

🔎 Evaluation

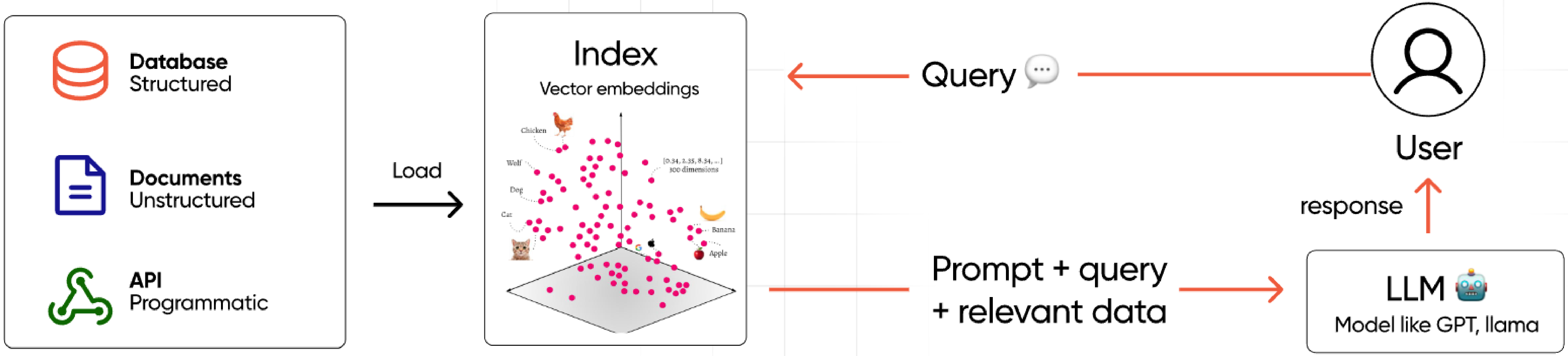
Steps

1) RAG → DATA Ingestion, Retriever, Generation ✓

2) TEST DATA ⇒ Question ↔ Answer

3) Evaluation Metrics [LLM as a judge]

Your Data



Database
Structured

Documents
Unstructured

API
Programmatic

Load

Index

Vector embeddings

Query 💬

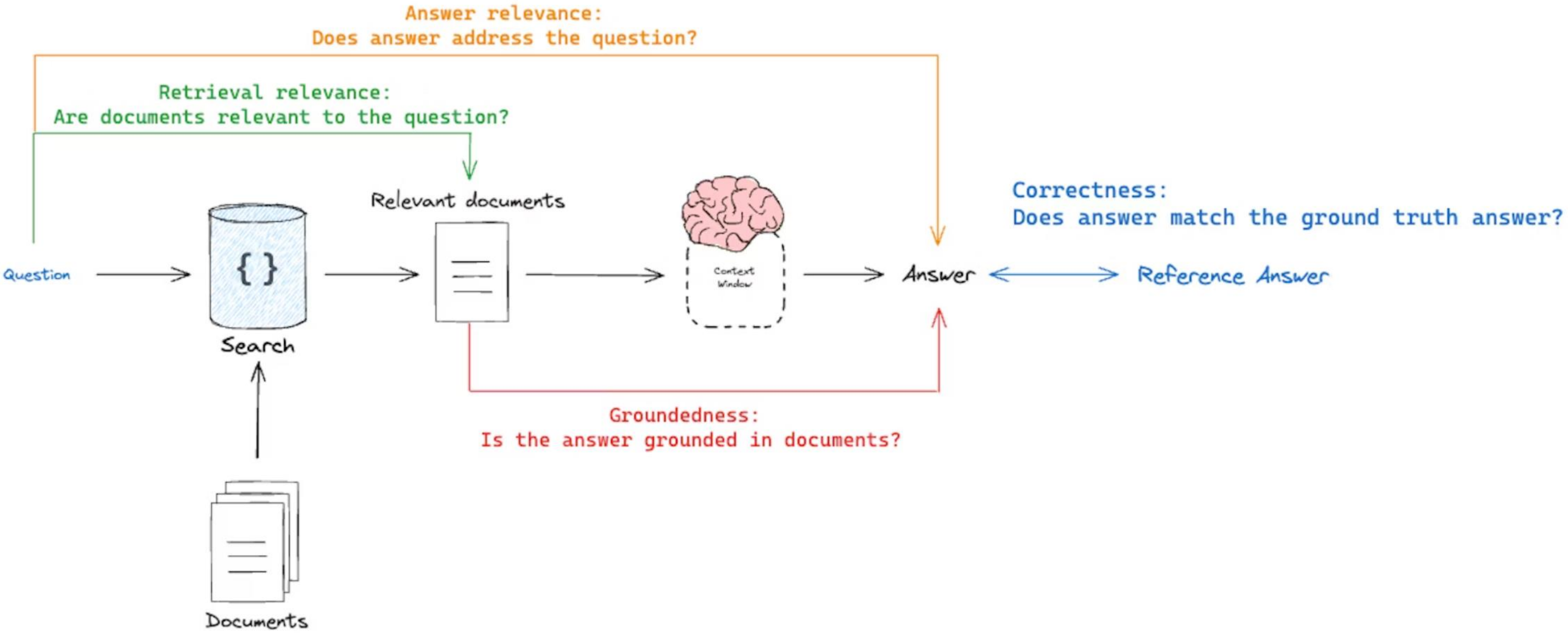
User

response

Prompt + query
+ relevant data

LLM 🤖

Model like GPT, llama



Answer relevance:
Does answer address the question?

Retrieval relevance:
Are documents relevant to the question?

Correctness:
Does answer match the ground truth answer?

Groundedness:
Is the answer grounded in documents?

Correctness: Response vs reference answer

- Goal: Measure “*how similar/correct is the RAG chain answer, relative to a ground-truth answer*”
- Mode: Requires a ground truth (reference) answer supplied through a dataset
- Evaluator: Use LLM-as-judge to assess answer correctness.

Relevance: Response vs input

- Goal: Measure "*how well does the generated response address the initial user input*"
- Mode: Does not require reference answer, because it will compare the answer to the input question
- Evaluator: Use LLM-as-judge to assess answer relevance, helpfulness, etc.

Groundedness: Response vs retrieved docs

- Goal: Measure “*to what extent does the generated response agree with the retrieved context*”
- Mode: Does not require reference answer, because it will compare the answer to the retrieved context
- Evaluator: Use LLM-as-judge to assess faithfulness, hallucinations, etc.

Retrieval relevance: Retrieved docs vs input

- Goal: Measure "*how relevant are my retrieved results for this query*"
- Mode: Does not require reference answer, because it will compare the question to the retrieved context
- Evaluator: Use LLM-as-judge to assess relevance