databricks

Evaluation

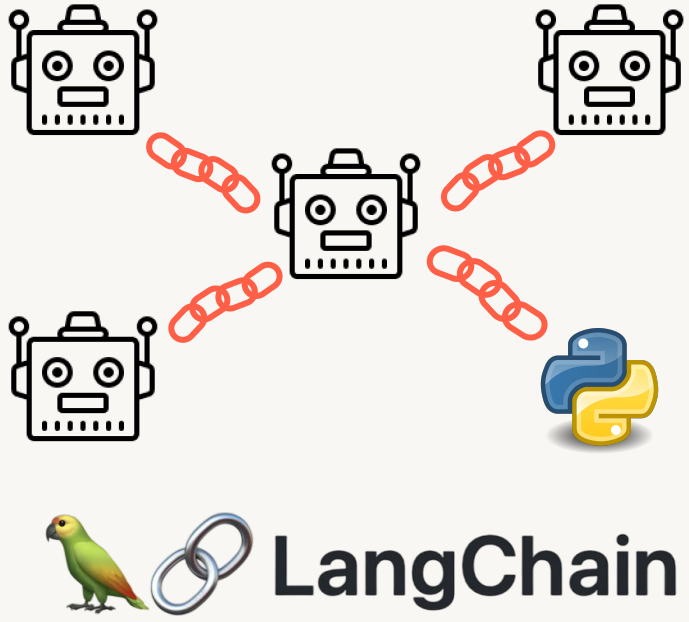
Of LLM Chains and Agents

Harrison Chase

Co-Founder and CEO at LangChain

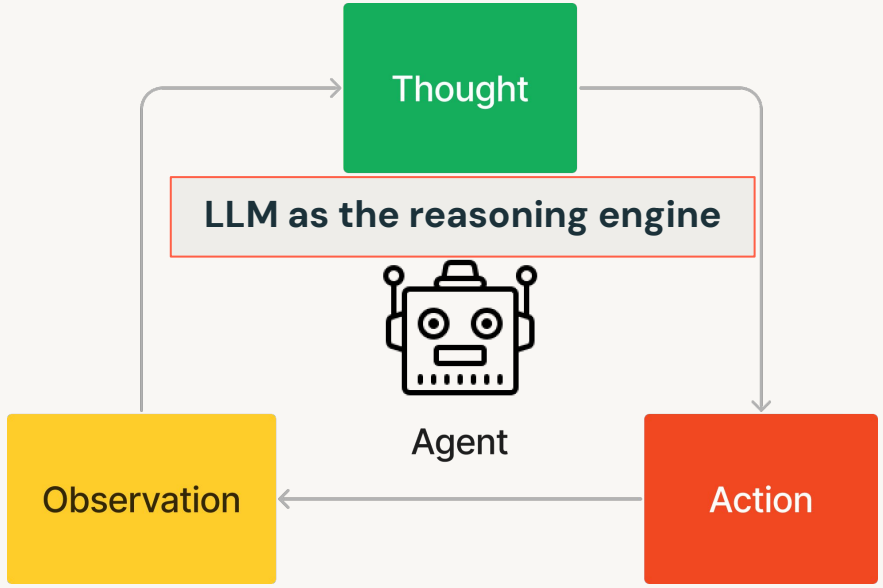
Agenda

- Overview of LLM chains and agents
- Why is evaluation hard?
  - Lack of data
  - Lack of metrics
- Potential solutions
- Offline evaluation
- Online evaluation

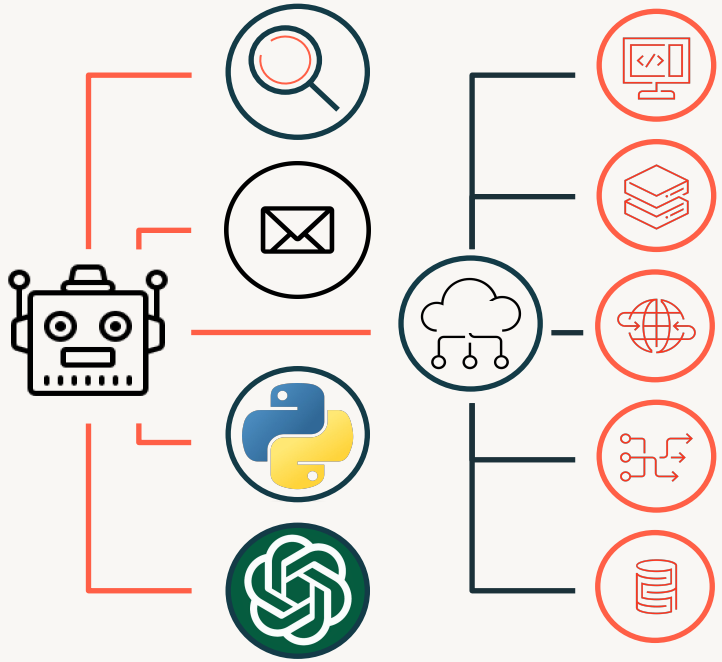


Overview of LLM Chains and Agents

LLM Agent

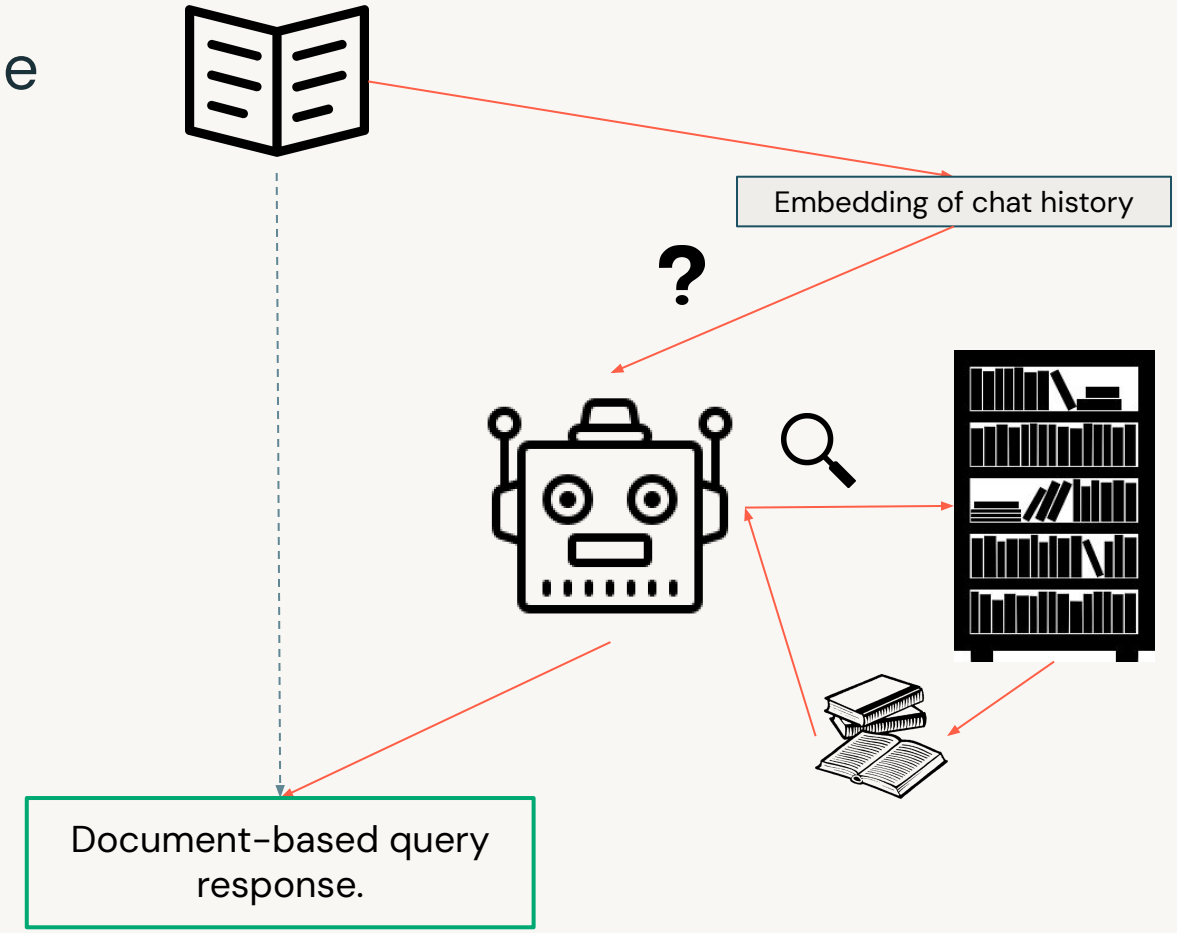


Connect it other sources of data and computation



Retrieval Augmented Generation Chatbot

1. Combine chat history into standalone question
2. Retrieve relevant documents
3. Generate final answer



Why is evaluation hard: Lack of data

Usually don't start with a dataset (cold start)

Instead with an idea or problem

Unclear what the dataset would even be (constantly changing)

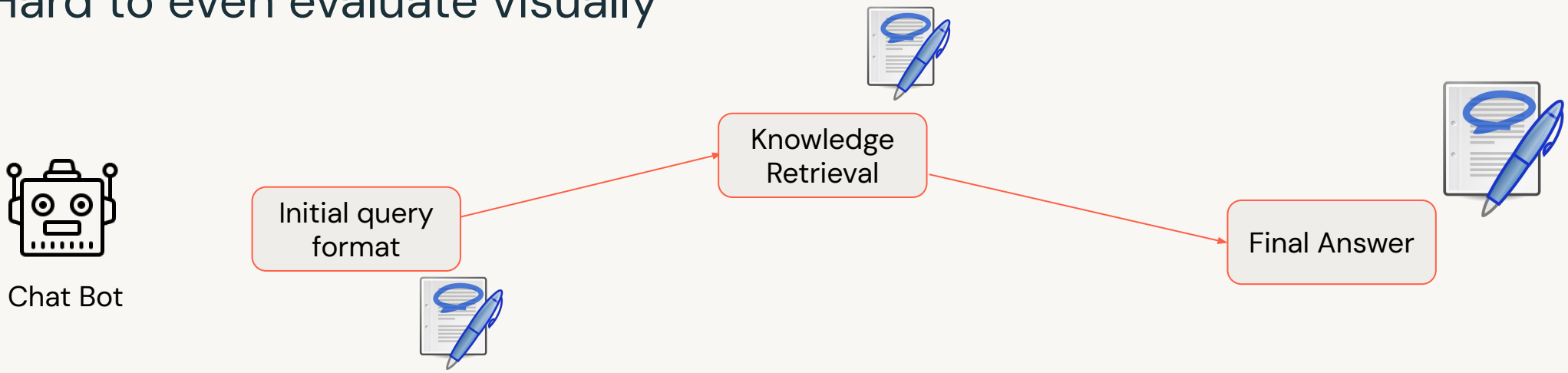
No ground truth to guide gathering data

Generic Data → **Easy** to evaluate

Specific/Scarce Data → **Hard** to evaluate

Why is evaluation hard: Lack of metrics

Hard to even evaluate visually



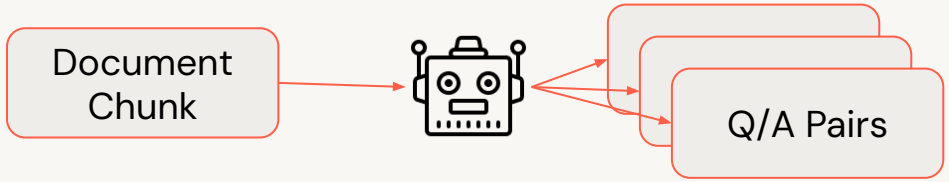
Traditional ML metrics don't work well

- Perplexity
- BLEU
- ROUGE
- SQuAD

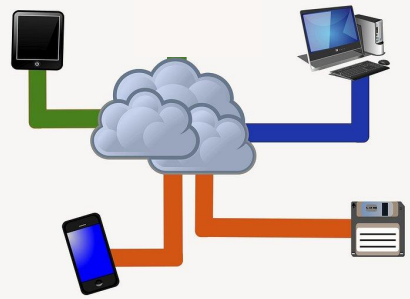
Potential Solutions – Best Practices

Lack of data

- Generate datasets using language model + chaining ahead of time:
  - An LLM can generate data from document chunks to be used as a test dataset.



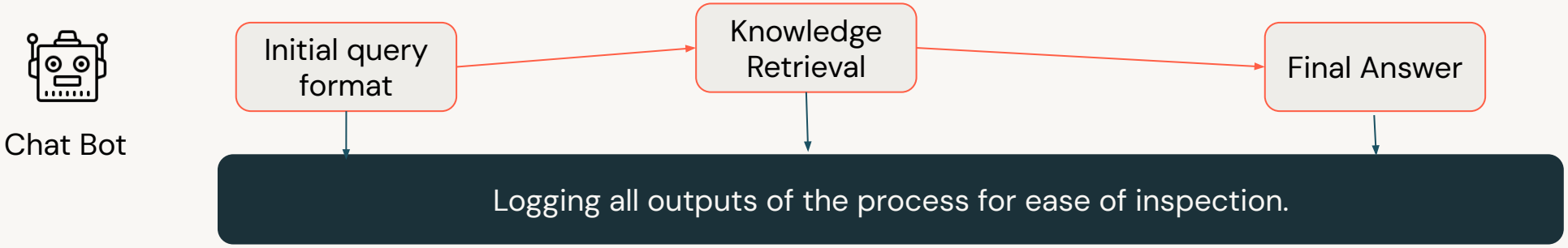
- Accumulate over time
  - Track I/O over time in production.



Potential Solutions – Best Practices

Lack of metrics

- Make it easy to inspect visually



- Use LLM as a judge.
  - Final answer judged by an LLM for semantic equivalence.

- User feedback, directly or indirectly.

Offline Evaluation

Procedure:

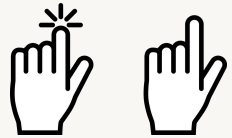
1. Create dataset of test data points to run against
2. Run chain or agent against them
3. Visually inspect them
4. Use LLM to auto-grade them

Online Evaluation

Gather feedback on each incoming datapoint:

- Direct feedback (thumbs up/thumbs down)

- Indirect feedback (clicked on link, did not click)



- Track feedback overtime

The Future of Evaluation in LLM Chains

- Still a very new area of applied research.
- Only now are applications coming online.
- Best practices will continue to emerge.

Thank you!