LangChain

Best practices for evaluating LLM applications

Ankush Gola, Co-Founder (ankush@langchain.dev)

Team + Company Overview

Nov 2022

First commit to LangChain OSS by co-founders Harrison Chase & Ankush Gola

Jan - April 2023

Incorporated the company

Raised from Benchmark & Sequoia

Present

Team of 27

Mostly engineers from Facebook, dbt Labs, Databricks, Ramp, Uber, Robust Intelligence, etc.

Momentum and Scale of the LangChain Ecosystem

85k+ stars on GitHub ♦ **2500+** contributors

80+ models supported ♦ **600+** tool integrations

35k+ applications live

14M downloads/month

Rakuten infor Moody's Morningstar Brex AXA

instacart ally zapier elastic BCG adyen

The promise of LLM-powered applications is extraordinary. Companies are going on a similar, frenzied journey of exploring how to harness the power of gen AI for their business.

1

What should we build? 🤔

Lots of options. Companies want to take a crawl, walk, run approach.

2

How do we go from curiosity to proof of concept? 🪄

Need a new toolchain and shift in strategy of building.

3

How do we meet our quality standards? 🛠️

Repeatability of results at scale requires a new way of testing and evaluation.

4

How do we build trust in a non-deterministic application? 🤝

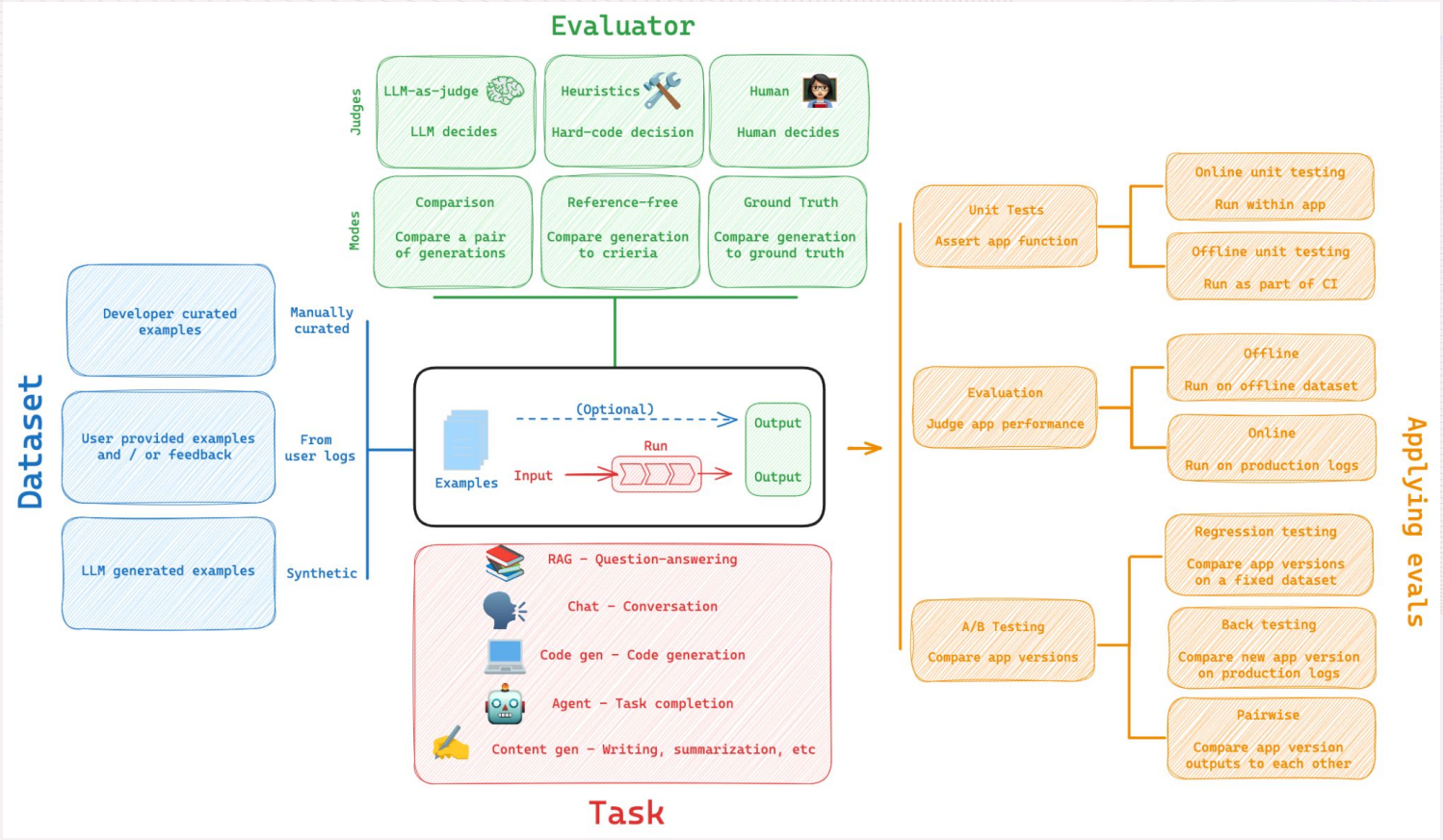
Minimize reputational risk of delivering poor experiences.

5

How do we keep customer data safe? 🔍

General uneasiness with how data is being stored and used.

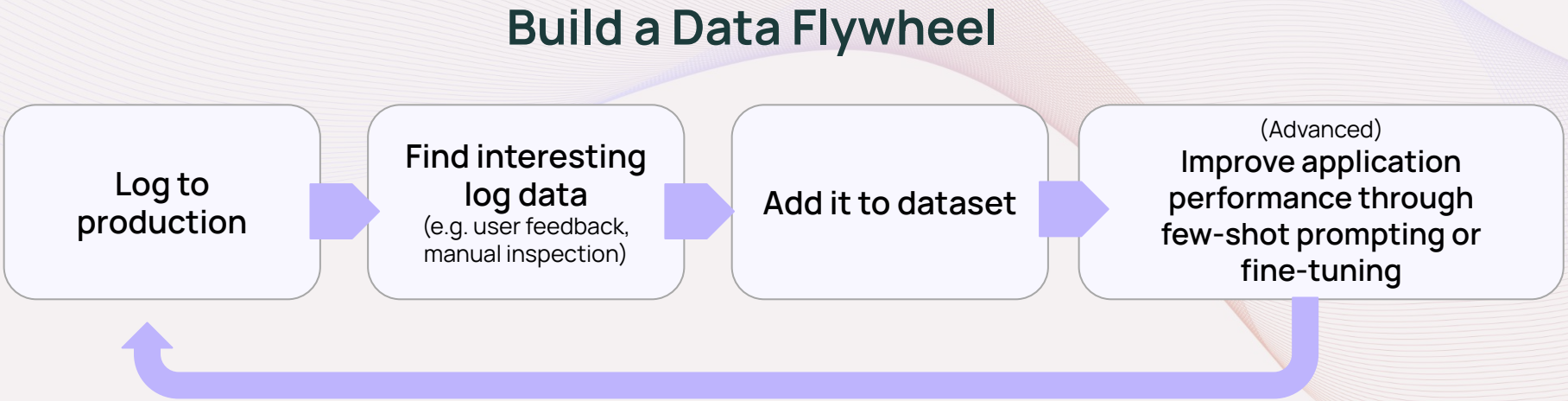
Evaluations: Measure the Performance of Your LLM App



Best Practices

1. **Datasets:** Add Interesting Production Logs to Datasets
2. **Evaluators:** Using LLM-as-a-Judge Evaluators
3. **Evaluators:** Evaluate on Intermediate Steps
4. **Applying Evals:** Online Evaluations

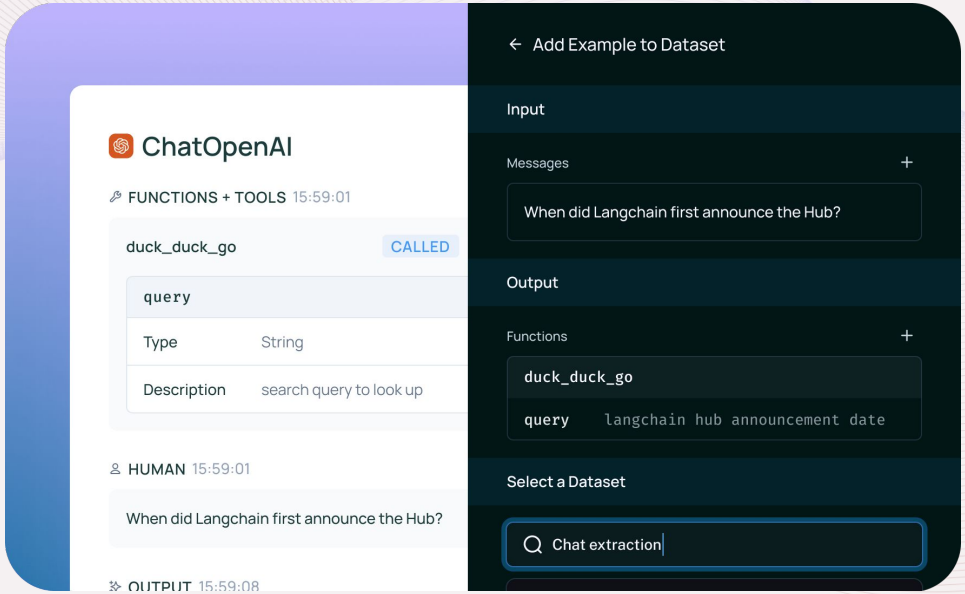
1 **Datasets:** Add Interesting Production Logs to Datasets



1 **Datasets:** Add Interesting Production Logs to Datasets (cont.)

- You can improve application performance with user feedback
- Dosu, the GitHub bot that auto-labels issues and PRs, uses LangSmith to power their in-context (few-shot) learning pipeline
- They used LangSmith to achieve **30% accuracy improvement with no prompt engineering** [Blog]

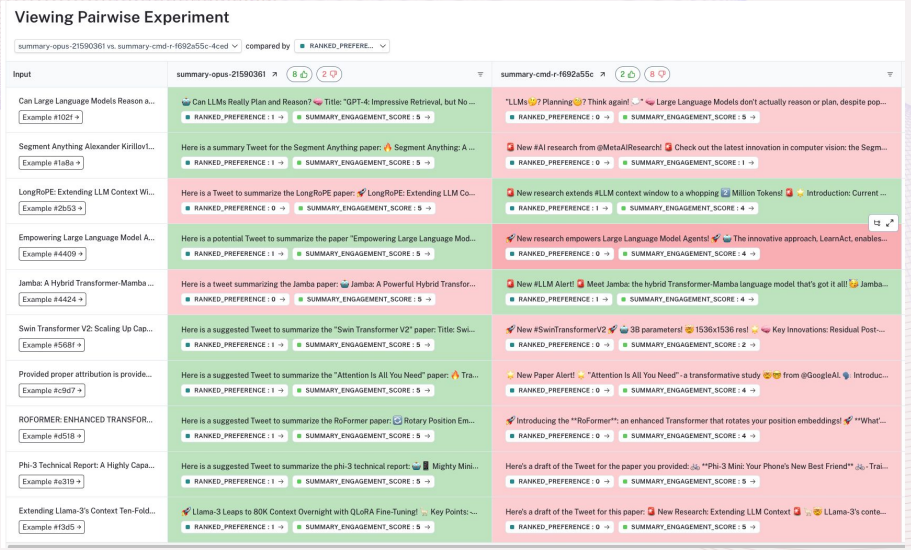
Demo



2 **Evaluators:** Using LLM-as-a-Judge Evaluators

- Prompt an LLM to score your LLM application results
- Use the most advanced model you can afford
  - Different tradeoff between speed and reasoning capability for evals
- Can be another source of noise
  - Use **pairwise evaluations** — theory that LLMs are better at ranking outputs than giving absolute score [Demo] [Example evaluator]
  - Audit evaluation results, measure LLM to human alignment

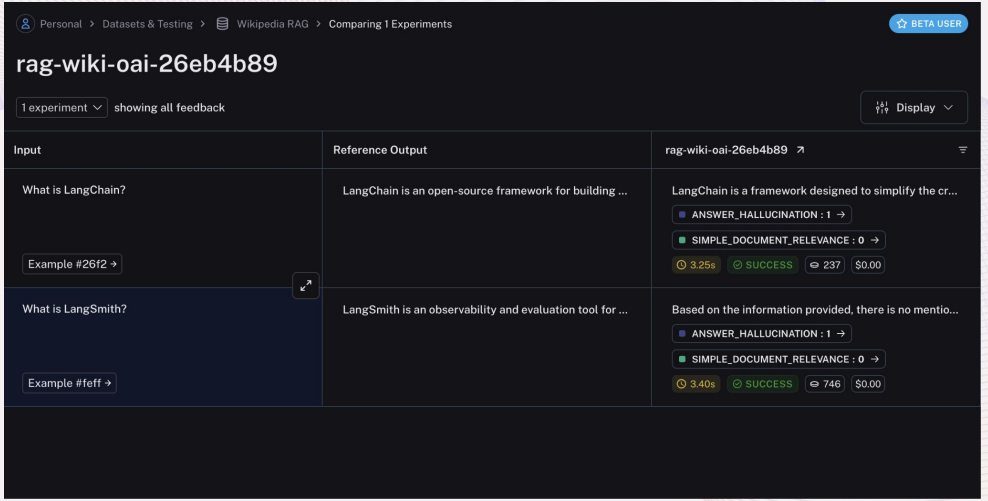
Docs



3 **Evaluators:** Evaluate on Intermediate Steps

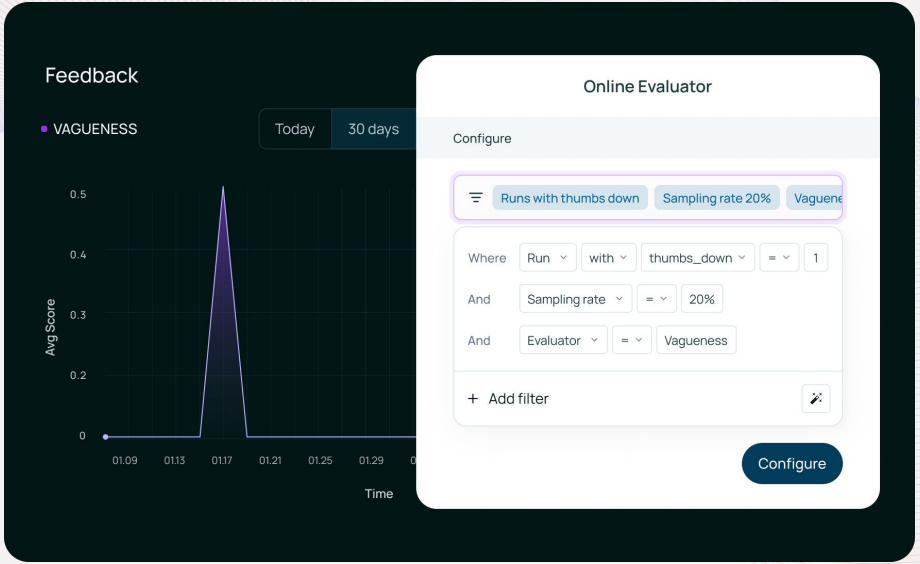
- Oftentimes, **evaluating the performance of intermediate steps of an LLM pipeline is just as important** as evaluating the final output
- **Especially relevant for RAG**, where you want to score on document relevance (eval retrieved docs wrt query) and hallucination (eval final generation wrt retrieved docs)
    - Also relevant agents

Demo **Docs**



4 **Applying Evals:** Online Evaluations

- **Useful to gather more signal in production** - A/B test multiple prompts or model choices on real-world scenarios
- Apply an automatic evaluator to a sample of production traces that pass a filter



Demo

Docs

Roadmap

https://docs.smith.langchain.com/

- Audit LLM results
- Better support for running evals in CI
- Trials
- Online evaluation for RAG

https://python.langchain.com/v0.1/docs

- Stability, integrations, security

https://langchain-ai.github.io/langgraph

- The best way to build language model agents
- Already adopted by many of the best companies building agents

Questions?

ankush@langchain.dev