

# LangChain

Best practices for evaluating LLM applications

Ankush Gola, Co-Founder ([ankush@langchain.dev](mailto:ankush@langchain.dev))

## Team + Company Overview

## Momentum and Scale of the LangChain Ecosystem



**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.

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

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

**35k+** applications live  
**14M** downloads/month





















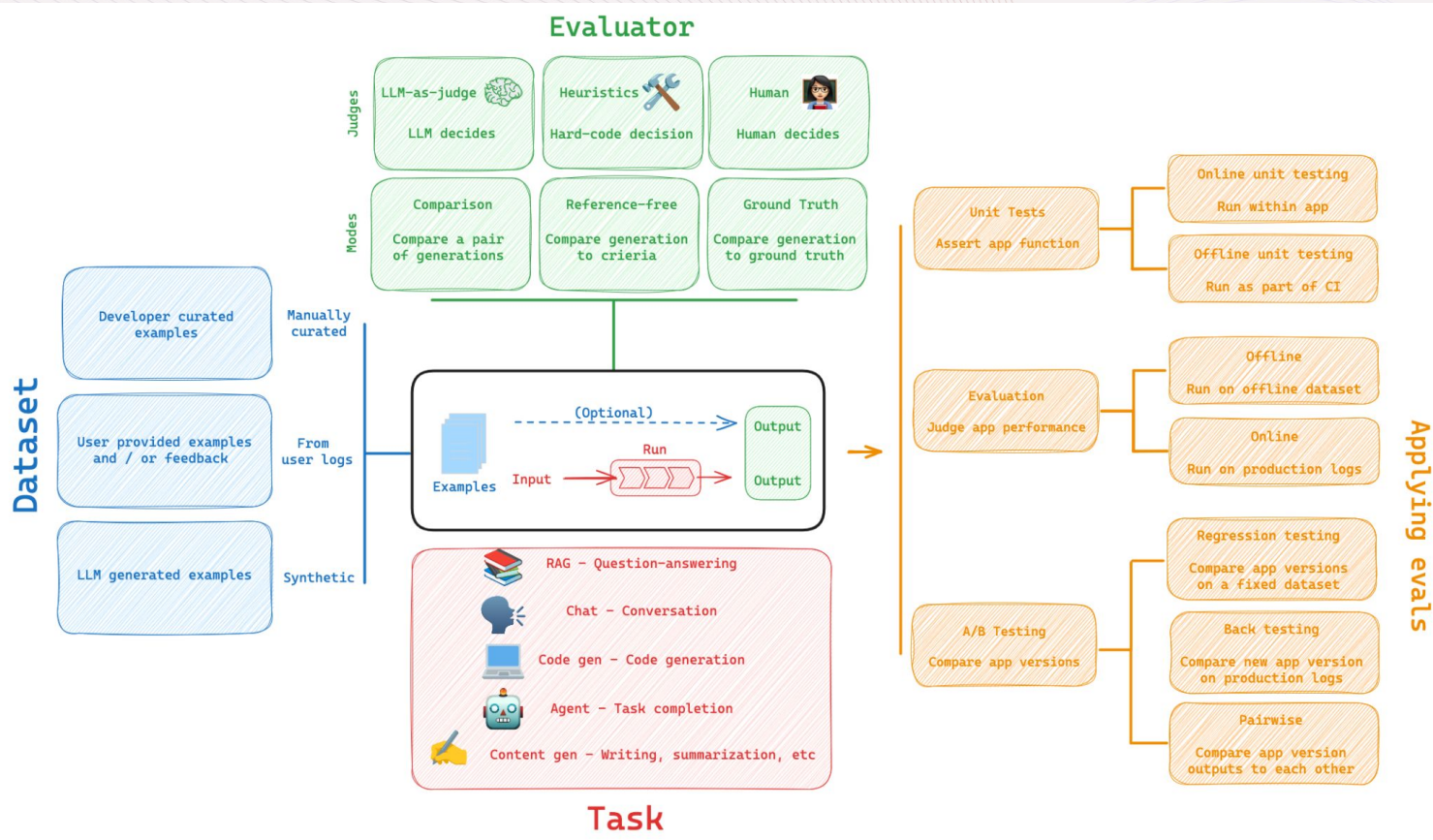




# The promise of LLM-powered applications is extraordinary. Companies are going on a similar, frenzied journey of exploring how to harness the power of gen AI for their business.

- 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

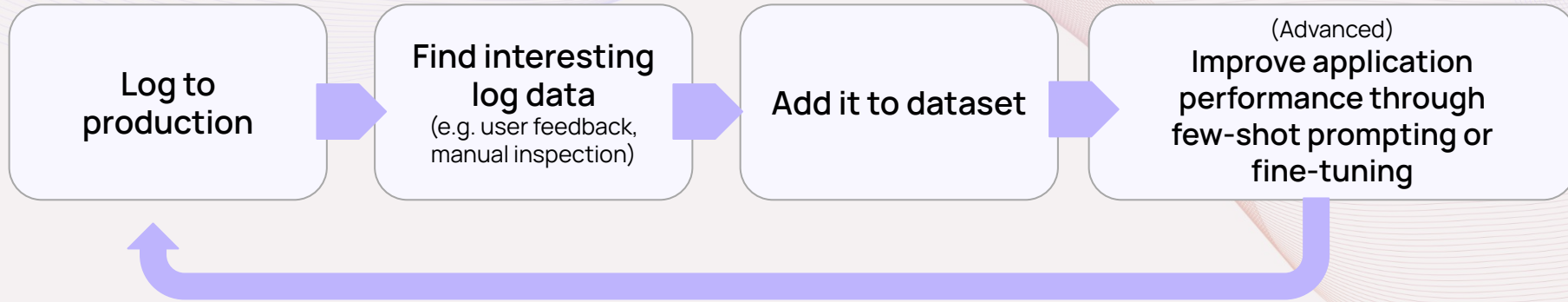


# Best Practices

- 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

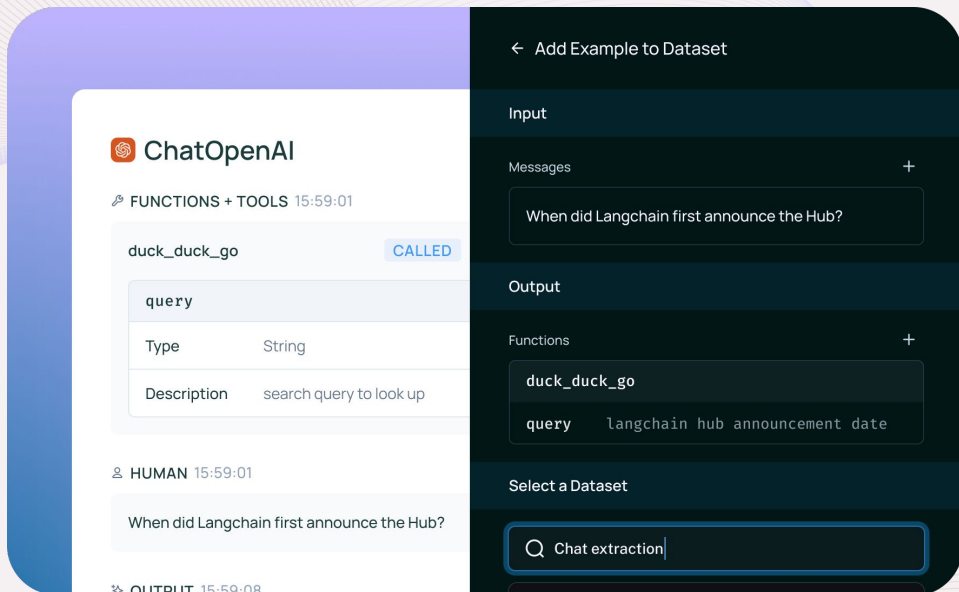
## Build a Data Flywheel



# 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

The screenshot displays a 'Viewing Pairwise Experiment' interface. At the top, it shows the comparison between 'summary-opus-21590361' and 'summary-cmd-r-692a55c-4ced', with a dropdown menu set to 'RANKED\_PREFER...'. Below this, a table lists various prompts and their corresponding outputs from two different models. Each row includes a 'Input' column, a 'summary-opus-21590361' column, and a 'summary-cmd-r-692a55c-4ced' column. The outputs are color-coded: green for higher scores and red for lower scores. Each output includes a 'RANKED\_PREFERENCE' score and a 'SUMMARY\_ENGAGEMENT\_SCORE'. The prompts cover a range of topics, including language model capabilities, summarization, and technical reports.

Input	summary-opus-21590361	summary-cmd-r-692a55c-4ced
Can Large Language Models Reason a...	Can LLMs Really Plan and Reason? Title: "GPT-4: Impressive Retrieval, but No ..." RANKED_PREFERENCE: 1 → SUMMARY_ENGAGEMENT_SCORE: 5 →	"LLMs? Planning? Think again!" Large Language Models don't actually reason or plan, despite pop-... RANKED_PREFERENCE: 0 → SUMMARY_ENGAGEMENT_SCORE: 5 →
Segment Anything Alexander Kirillov...	Here is a summary Tweet for the Segment Anything paper: Segment Anything: A ... RANKED_PREFERENCE: 1 → SUMMARY_ENGAGEMENT_SCORE: 5 →	New AI research from @MetaAIResearch! Check out the latest innovation in computer vision: the Segm-... RANKED_PREFERENCE: 0 → SUMMARY_ENGAGEMENT_SCORE: 1 →
LongRoPE: Extending LLM Context Wi...	Here is a Tweet to summarize the LongRoPE paper: LongRoPE: Extending LLM Co-... RANKED_PREFERENCE: 0 → SUMMARY_ENGAGEMENT_SCORE: 5 →	New research extends #LLM context window to a whopping 1 Million Tokens! Introduction: Current ... RANKED_PREFERENCE: 1 → SUMMARY_ENGAGEMENT_SCORE: 4 →
Empowering Large Language Model A...	Here is a potential Tweet to summarize the paper "Empowering Large Language Mod-... RANKED_PREFERENCE: 1 → SUMMARY_ENGAGEMENT_SCORE: 5 →	New research empowers Large Language Model Agents! The innovative approach, LearnAct, enables... RANKED_PREFERENCE: 0 → SUMMARY_ENGAGEMENT_SCORE: 4 →
Jamba: A Hybrid Transformer-Mamba ...	Here is a tweet summarizing the Jamba paper: Jamba: A Powerful Hybrid Transfor-... RANKED_PREFERENCE: 0 → SUMMARY_ENGAGEMENT_SCORE: 5 →	New #LLM Alert! Meet Jamba: the hybrid Transformer-Mamba language model that's got it all! Jamba-... RANKED_PREFERENCE: 1 → SUMMARY_ENGAGEMENT_SCORE: 4 →
Swin Transformer V2: Scaling Up Cap...	Here is a suggested Tweet to summarize the "Swin Transformer V2" paper: Title: Swi-... RANKED_PREFERENCE: 1 → SUMMARY_ENGAGEMENT_SCORE: 5 →	New #SwinTransformerV2! 38 parameters! 1536x1536 res! Key Innovations: Residual Post-... RANKED_PREFERENCE: 0 → SUMMARY_ENGAGEMENT_SCORE: 2 →
Provided proper attribution is provide...	Here is a suggested Tweet to summarize the "Attention Is All You Need" paper: Tra-... RANKED_PREFERENCE: 1 → SUMMARY_ENGAGEMENT_SCORE: 5 →	New Paper Alert! "Attention Is All You Need" a transformative study from @GoogleAI. Introduc-... RANKED_PREFERENCE: 0 → SUMMARY_ENGAGEMENT_SCORE: 5 →
ROFORMER: ENHANCED TRANSFORMER...	Here is a suggested Tweet to summarize the RoFormer paper: Rotary Position Em-... RANKED_PREFERENCE: 1 → SUMMARY_ENGAGEMENT_SCORE: 5 →	Introducing the "RoFormer": an enhanced Transformer that rotates your position embeddings! "What"... RANKED_PREFERENCE: 0 → SUMMARY_ENGAGEMENT_SCORE: 4 →
Phi-3 Technical Report: A Highly Capa...	Here is a suggested Tweet to summarize the phi-3 technical report: Mighty Mini-... RANKED_PREFERENCE: 1 → SUMMARY_ENGAGEMENT_SCORE: 5 →	Here's a draft of the Tweet for the paper you provided: "Phi-3 Mini: Your Phone's New Best Friend" Trai-... RANKED_PREFERENCE: 0 → SUMMARY_ENGAGEMENT_SCORE: 5 →
Extending Llama-3's Context Ten-Fold...	Llama-3 Leaps to 80K Context Overnight with QLoRA Fine-Tuning! Key Points: ... RANKED_PREFERENCE: 1 → SUMMARY_ENGAGEMENT_SCORE: 5 →	Here's a draft of the Tweet for this paper: New Research: Extending LLM Context Llama-3's conte-... RANKED_PREFERENCE: 0 → SUMMARY_ENGAGEMENT_SCORE: 5 →

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

The screenshot shows a dashboard for an experiment named 'rag-wiki-oai-26eb4b89'. It features a table with three columns: 'Input', 'Reference Output', and a results column. The results column displays metrics for 'ANSWER\_HALLUCINATION' and 'SIMPLE\_DOCUMENT\_RELEVANCE' for two different input queries. The first query, 'What is LangChain?', has a score of 3.25s, is marked as SUCCESS, and has 237 evaluations with a cost of \$0.00. The second query, 'What is LangSmith?', has a score of 3.40s, is marked as SUCCESS, and has 746 evaluations with a cost of \$0.00.

Input	Reference Output	rag-wiki-oai-26eb4b89
What is LangChain? Example #26f2 →	LangChain is an open-source framework for building ...	LangChain is a framework designed to simplify the cr... ANSWER_HALLUCINATION : 1 → SIMPLE_DOCUMENT_RELEVANCE : 0 → 3.25s SUCCESS 237 \$0.00
What is LangSmith? Example #feff →	LangSmith is an observability and evaluation tool for ...	Based on the information provided, there is no mentio... ANSWER_HALLUCINATION : 1 → SIMPLE_DOCUMENT_RELEVANCE : 0 → 3.40s SUCCESS 746 \$0.00

Demo

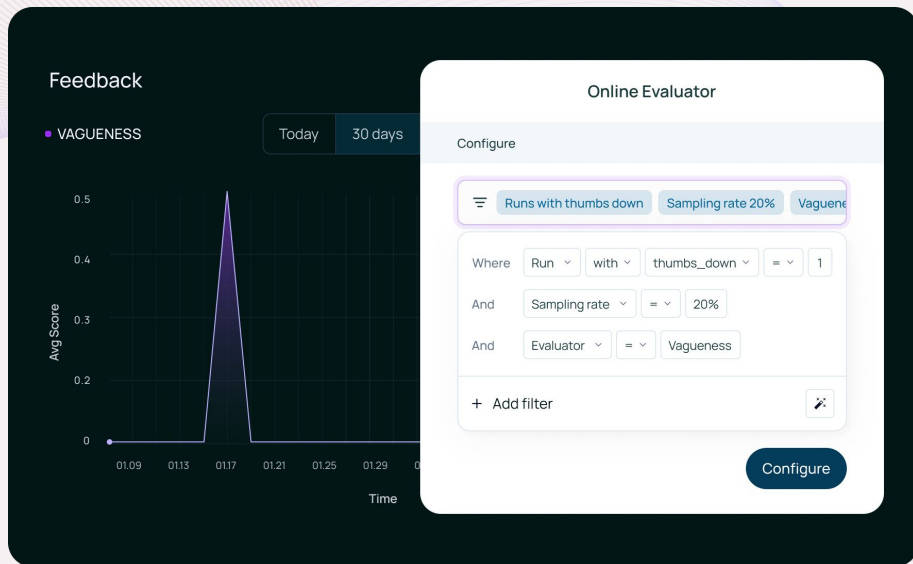
Docs

## 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



**LangSmith**

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

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



**LangChain**

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

- Stability, integrations, security



**LangGraph**

<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](mailto:ankush@langchain.dev)**