

# Uber's GenAI Oncall Co-Pilot Journey

Paarth Chothani, Staff Software Eng, Uber

Nicholas Marcott, Staff Software Eng TLM, Uber

Eduards Sidorovics, Sr Software Eng, Uber



FEATURE STORE SUMMIT 2024

**DATA FOR AI:**  
REAL-TIME, BATCH, AND LLMs

Organized by  **HOPSWORKS**

# Day of an Oncall



# Cumbersome Process





# Hard To Find Information



# On-Call Copilot Challenges

- Response Accuracy
- Data Quality
- Data Security
- UX Experience
- Bot Credibility
- Human Bias

# Solution Overview: Genie



**Eduards Sidorovics** Just now

!q What is Michelangelo?

2 replies



**BMO APP** Just now

Genie 🧙 - Gen AI Oncall Copilot Please wait for an 🙋 on-call engineer to assist you.

Michelangelo is Uber's internal end-to-end machine learning platform built by the AI platform team. It enables ML developers to prepare data, train and evaluate models, deploy models and make predictions (both batch and online), set up auto-retrain pipelines, and monitor model performance in production. Its mission is to provide Uber's ML developers with advanced ML capabilities and tools to rapidly build, deploy, and iterate high-quality ML applications at scale.

Additional References

[\[1\]](#) [\[2\]](#) [\[3\]](#) [\[4\]](#)

✅ Resolved By Genie

👍 Helpful

👎 Not Helpful

✖ Not Relevant

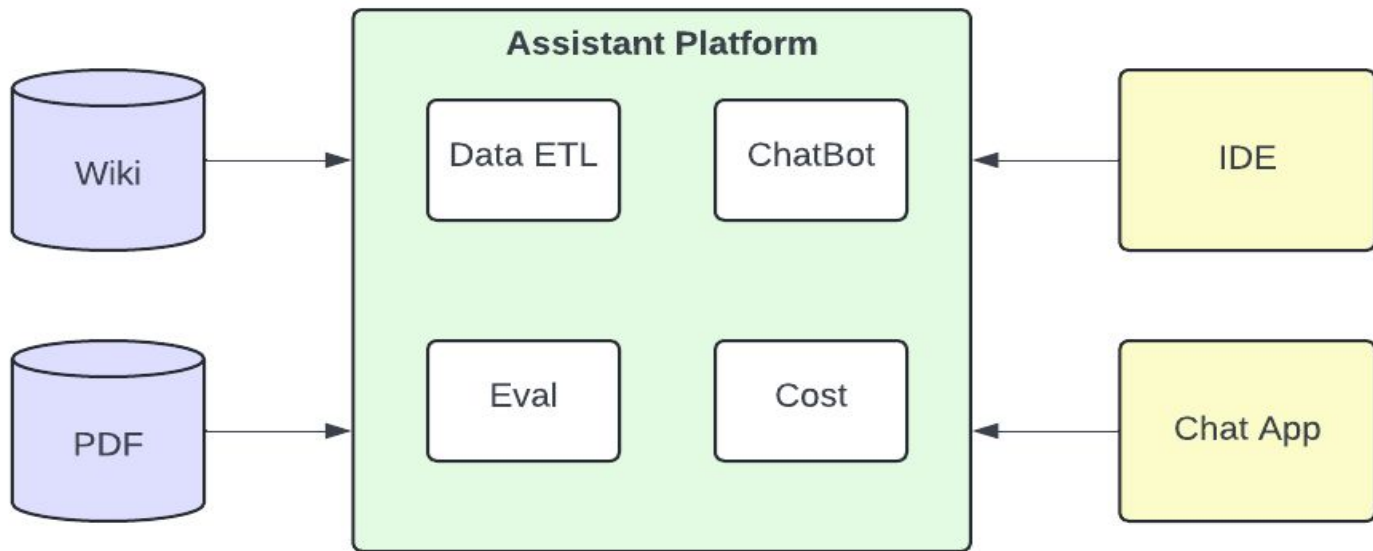
## What is Genie?

- A Slack chatbot that utilizes GenAI
- Deployed within Slack channels for immediate user support.

## Objectives

- Improve productivity for users and on-call engineers.
- Provide accurate, relevant, and secure information.
- Collect user feedback

# Architecture



# Michelangelo Studio Setup

Michelangelo Studio

Token Docs Help Report bugs

Home > ma-genai-customer-sandbox > Gen AI > Prepare Data > Pipelines

### Prepare Data

Create RAG, eval pipelines and upload datasets.

Pipelines Runs Data Sources Evaluation Reports

Search... Pipelines Revisions Add filter Create pipeline

Name ↑	Last updated ↓	Type ↑	Owner ↑	Branch ↑	St
genie-embedding-ingestion-hive Revision 98036f51c69c	2024/05/15 22:36:55	Embedding Generation	Jin Sun jins	master	
genai-customer-sandbox-pdf-ingestion Revision 98036f51c69c	2024/05/15 22:36:51	Embedding Generation	Jin Sun jins	master	
engwiki-ingestion-to-hive Revision 98036f51c69c	2024/05/15 22:36:43	Embedding Generation	Jin Sun jins	master	
genie-embedding-ingestion-hive Revision b8ea800bb64e	2024/05/09 03:17:34	Embedding Generation	Jonathan Li ljonath	master	
genai-customer-sandbox-pdf-ingestion Revision b8ea800bb64e	2024/05/09 03:17:30	Embedding Generation	Jonathan Li ljonath	master	
engwiki-ingestion-to-hive Revision b8ea800bb64e	2024/05/09 03:17:22	Embedding Generation	Jonathan Li ljonath	master	

Engwiki

Stack

Files

Source type \*

Engwiki

EngWiki spaces \*

GOMONREPO GOFRAMEWORK

For example: GOMONREPO, GOFRAMEWORK, etc.

Source type \*

Stack

StackOverflow tags \*

kafka go-monorepo

For example: kafka, go-monorepo, etc.

Source type \*

File

Files \*

Drop files here to upload...  
Supported types – pdf, csv, md, docx. Max size – 50MB

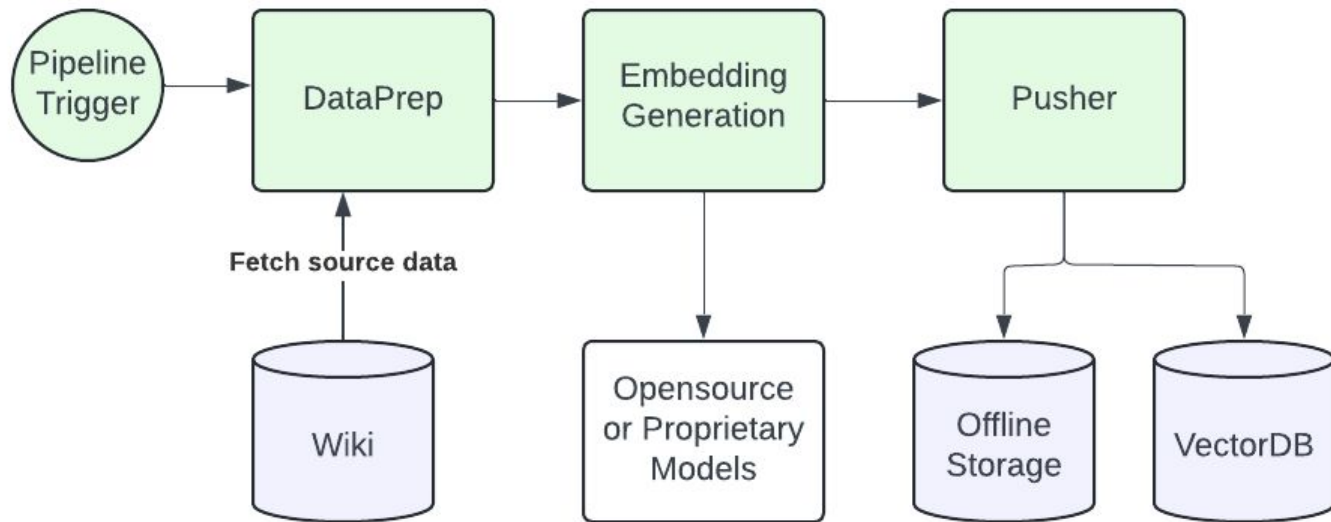
Browse files

/staging/michelangelo-studio-web/form-uploads/rag-pipeline/2024-05-23\_10-59-03\_361\_rag\_pdf\_1.pdf

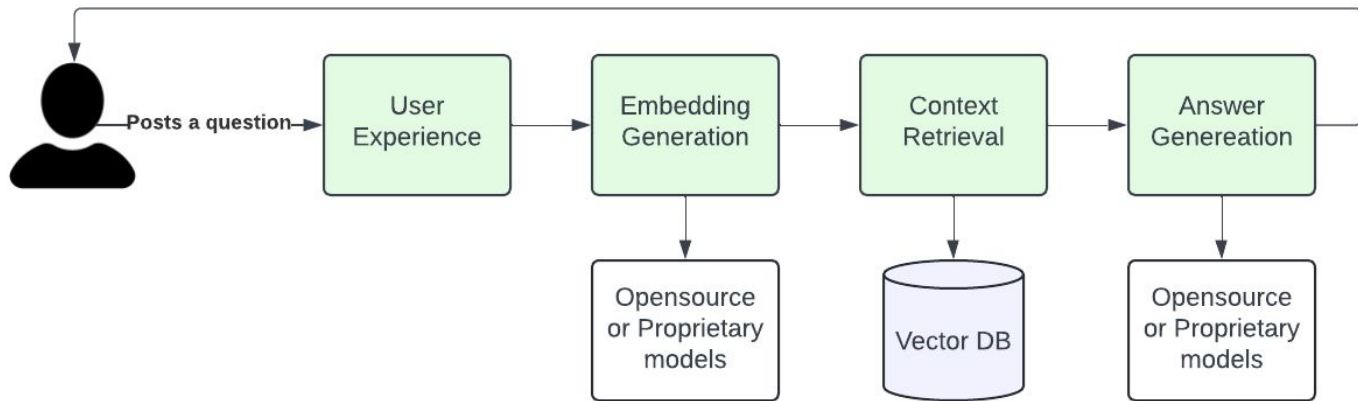
/staging/michelangelo-studio-web/form-uploads/rag-pipeline/2024-05-23\_10-59-04\_524\_rag\_pdf\_2.pdf



# Data ETL

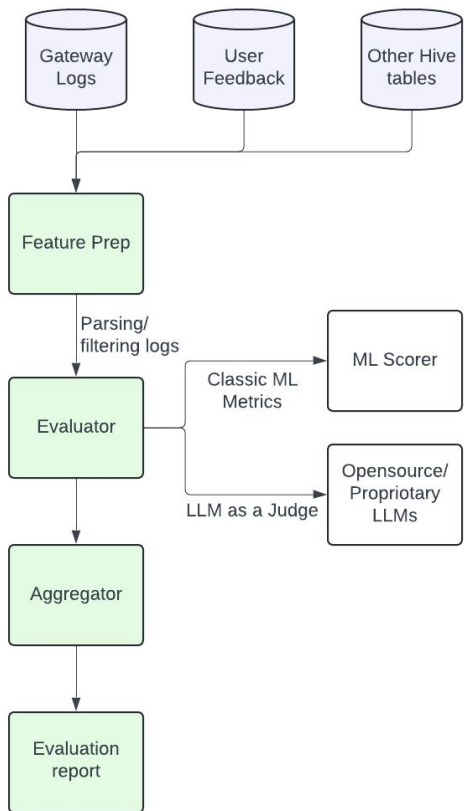


# Platform



- Cost tracking
- Overriding prompts
- Citation

# Performance Eval



## Input:

- LLM Response/Context
- User Feedback
- Other Metadata

## Feature Prep:

- Input processing and data denormalization

## Evaluator:

- LLM as a judge
- Foundational NLP metrics

## Aggregator:

- Calculate descriptive stats (i.e. mean, median)
- Summarize textual output

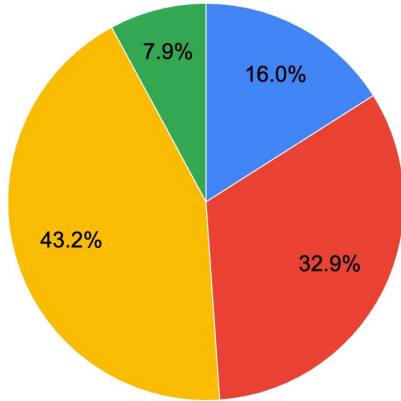
## Evaluation Report:

- Publish in UI for users

# Performance Eval - Demo



# Results/Conclusion



● Resolved by Genie ● Helpful ● Not Helpful ● Not Relevant

**154+** channels  
**> 70,000** question  
**48.9%** helpfulness rate

- We estimate It has saved us **13,000 engineering hours** so far since its launch.
- Productionizing a Bot is hard work
- Partner Engagement is critical to Bot accuracy
- Platformization led to self service RAG capabilities



# Thank you

[Blog](#)

<https://tinyurl.com/uber-genie>



FEATURE STORE SUMMIT 2024

**DATA FOR AI:**  
REAL-TIME, BATCH, AND LLMS

Organized by  **HOPSWORKS**