# AIPI Chatbot –
# Fine-tuned Mistral 7B & CRAG

- Mrinoy, Leslie, Kahlia, Samyukta

# Table of contents

01 Motivation

02 Dataset

03 Fine-Tuning

04 System Architecture

05 Inference

06 Evaluation

07 Cost

01

# Motivation

# Motivation

- High speed search and scalability
- Better integration
- Easy Hosting

- High Benchmark results
- Efficient Architecture
- Relatively lightweight

- **CRAG** based pipeline hooked with tavily search
- Directs questions with low match to web search
- Enhances bots ability to handle out-of-context questions.

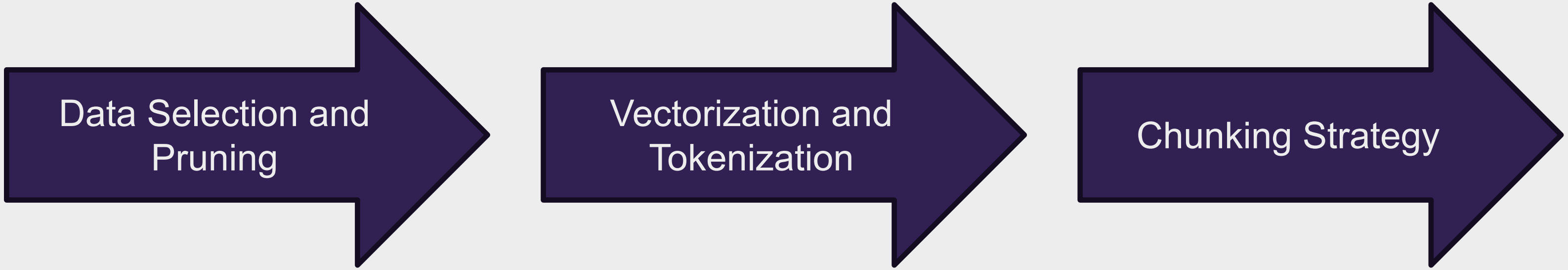
02

# Dataset

# Data

1. AIPI FAQ Document
2. Duke AIPI Web domain
   - https://ai.meng.duke.edu/
   - All subdomains attached to this master domain
3. Syllabus Information for AIPI courses
   - Excluding certain ones we did not have access to, i.e. AIPI 560

# Data Preprocessing



Data Selection and Pruning

Vectorization and Tokenization

Chunking Strategy

Editing files to remove:

- Information on Duke Medx, PhDs
- Remove old content that is irrelevant.
- Irrelevant headers
- Adding document context (syllabus)

WhereIsAI/UAE-Large-V1 tokenizer –

- Model Size (Million Parameters): 335
- Retrieval Average: 54.66
- Summarization Average: 32.03

- Data Integrity and Leakage Prevention
- Chunk Overlap – 50
- Chunk Size – 500 tokens maximum
- Create embeddings solely from within same documents

03

# Fine Tuning

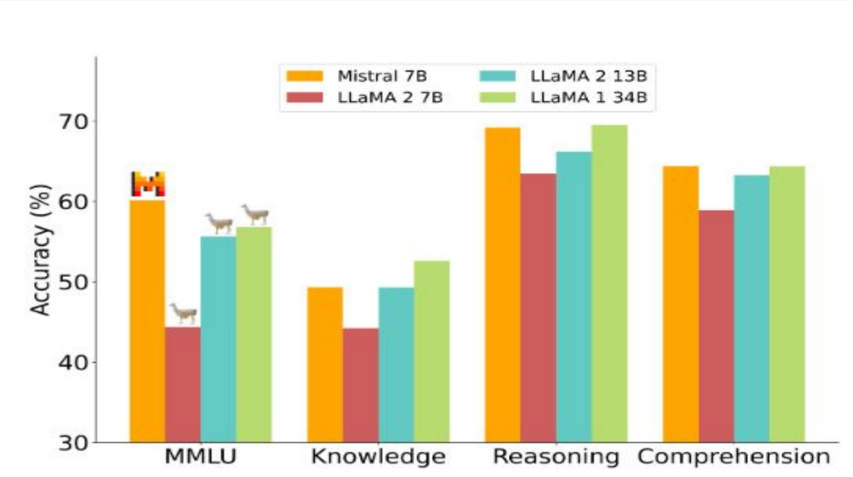
# Databricks Dolly 15K

01 **15,000 human-generated instruction corpus** specifically designed for training conversational AI

02 Can be used, modified, and extended for any purpose, including academic or commercial applications

**Training-test split: 80-20**

# Mistral-7B-v0.1



01 Efficient architecture with Grouped Query Attention (GQA) and Sliding Window Attention (SWA)

# Model Configurations

01

## ChatML format

- No Need for Instruct Tags
- Enhanced Format Clarity
- Optimized for RAG

02

## BitsAndBytes Quantized

- Models are loaded in 4-bit precision to decrease memory usage.
- Utilizes 'torch.bfloat16' for computing, balancing performance and precision.

03

## LoRA

Applies to all linear layers, enhancing the model's ability to adapt.

04

## Flash Attention

- Reduced Memory Footprint
- Increased Computational Speed
- Scalability

05

## Additional

- Optimized for 1 GPU - Training & Inference
- Optimizer – Adamw_bnb_8bit

# Results

## Training

- AWS g5.16xlarge
- GPU: Nvidia A10 (24GB)
- All experiments: 24 hrs
- Final model: 103 mins

## Training Costs

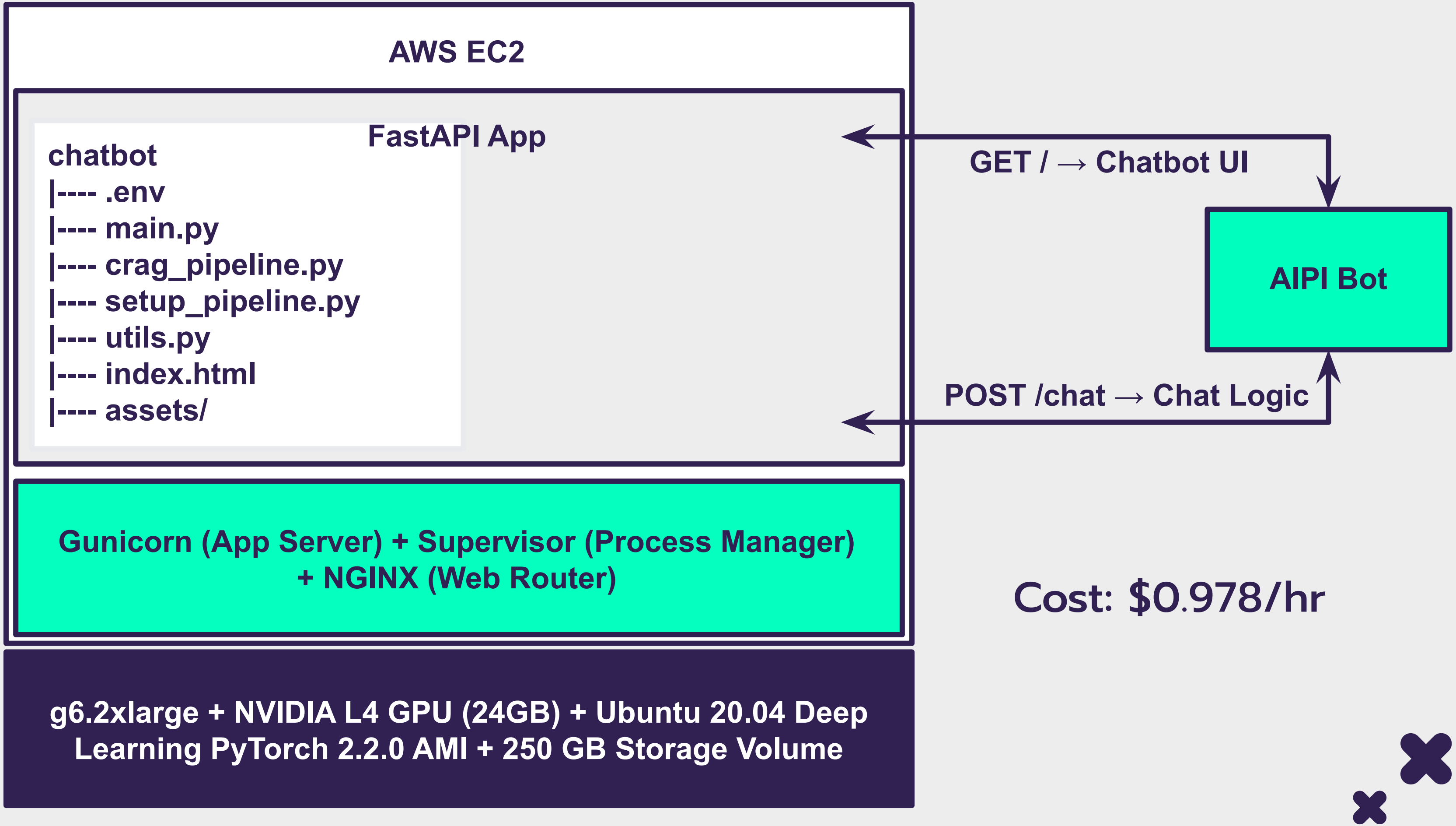
- All experiments: ~$100
- Final model: ~$7

## Evaluation

- 1000 random samples
- LLM as Judge – the LLaMa2-7B model serves as the standard, comparing the outputs generated by the Mistral bot against established ground truth data.
- Accuracy – 82.7%

04

# System Architecture



AWS EC2

FastAPI App

```
chatbot
|---- .env
|---- main.py
|---- crag_pipeline.py
|---- setup_pipeline.py
|---- utils.py
|---- index.html
|---- assets/
```

GET / → Chatbot UI

AIPI Bot

POST /chat → Chat Logic

Gunicorn (App Server) + Supervisor (Process Manager) + NGINX (Web Router)

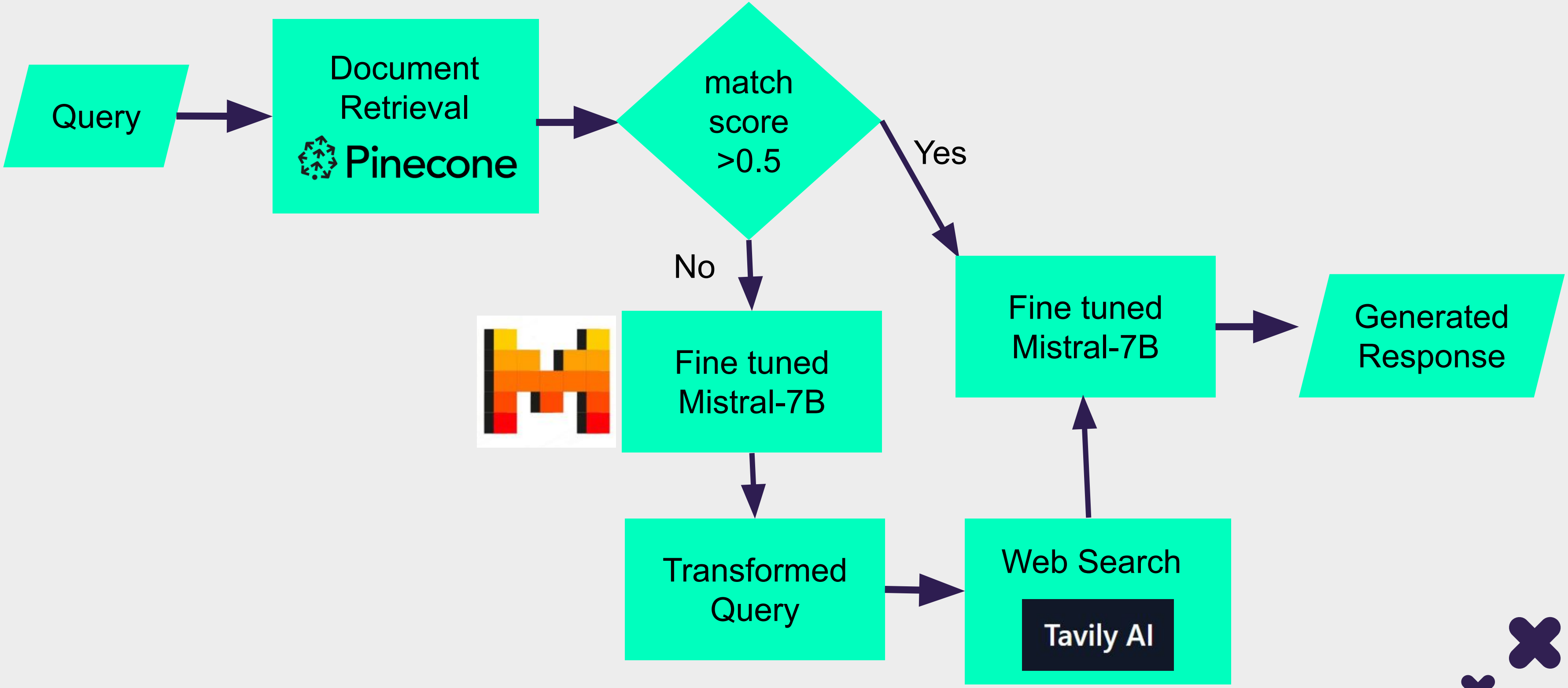
g6.2xlarge + NVIDIA L4 GPU (24GB) + Ubuntu 20.04 Deep Learning PyTorch 2.2.0 AMI + 250 GB Storage Volume

Cost: $0.978/hr

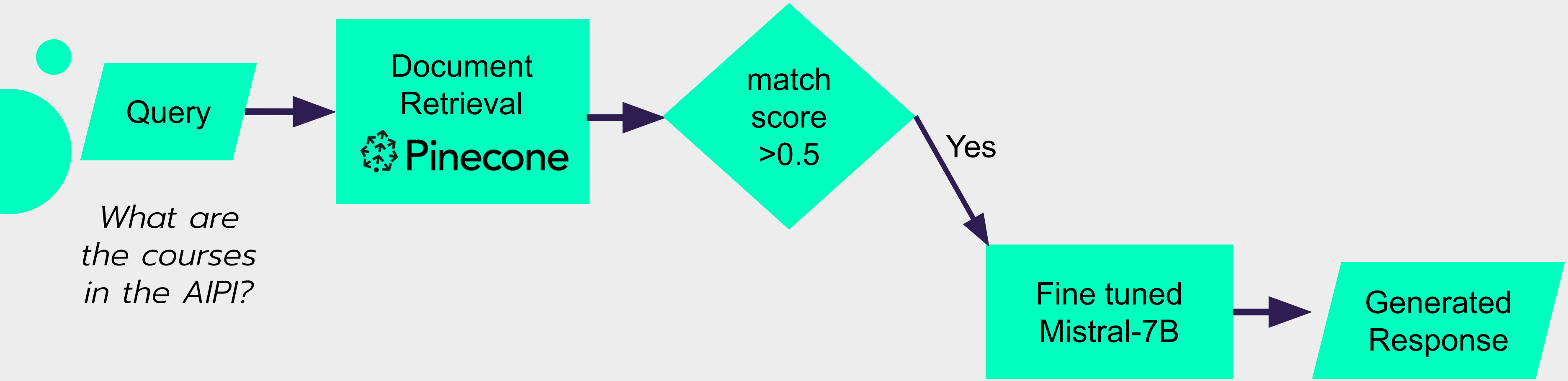
05

# Inference

# cRAG Pipeline



# cRAG Pipeline



*What are the courses in the AIPI?*

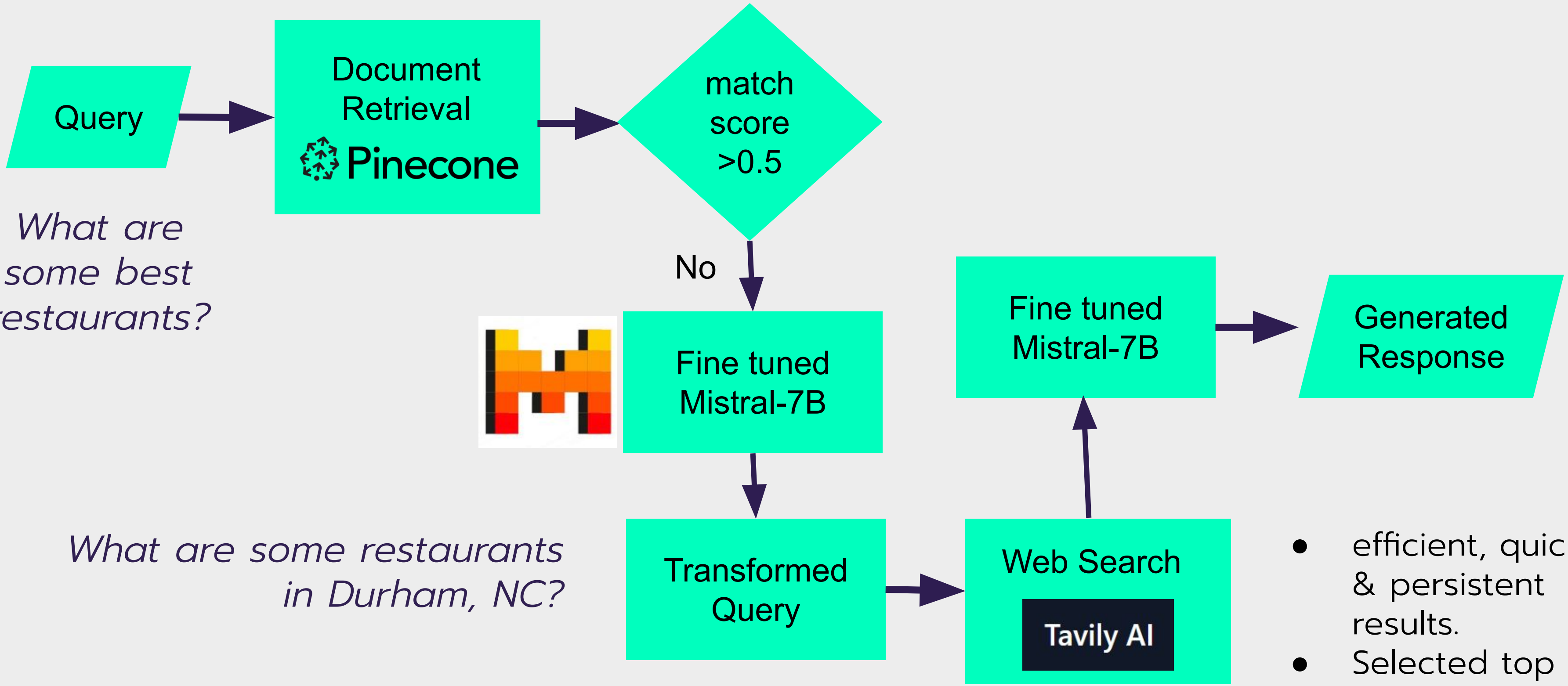
## Key Points

- Pinecone: Easy to Host
- Token size = 600
- Top 2 queries were selected

## Cost

- 10k queries/month
- 2k writes/month
- 10k vectors
- $0.45/month

# cRAG Pipeline



*What are some best restaurants?*

*What are some restaurants in Durham, NC?*

- efficient, quick & persistent results.
- Selected top 1

06

# Evaluation

# Results

Q/A

## Domain Questions

- 27 catered questions to Duke AIPI
- Subset of data about broad Duke information

## User Observations

- Response is based on content retrieval from Pinecone & web search
- AIPI specific questions perform better without hitting the web search
- Infrequent yet present hallucination

## Evaluation

- Human-as-a-Judge
- Rated on 1-5 scale
- Judged on information correctness & response formatting

- **AIPI Questions: 2.95**
- **Duke Questions: 2.2**

07

# Cost

# Pricing breakdown

| | |
|---|---|
| Experimentation: | ~$100 total |
| Model Fine-tuning: | ~$7 total |
| AWS Deployment: | ~$700/month |
| Pinecone DB: | ~$0.45/month |
| TOTAL | ~$8455/year |

# Thank you!